# Combining Crowd and Expert Labels Using Decision Theoretic Active Learning

**An T. Nguyen**
Department of Computer Science
University of Texas at Austin
atn@cs.utexas.edu

**Byron C. Wallace** and **Matthew Lease**
School of Information
University of Texas at Austin
{byron.wallace | ml}@utexas.edu

## Abstract

We consider a *finite-pool* data categorization scenario which requires exhaustively classifying a given set of examples with a limited budget. We adopt a hybrid human-machine approach which blends automatic machine learning with human labeling across a tiered workforce composed of domain experts and crowd workers. To effectively achieve high-accuracy labels over the instances in the pool at minimal cost, we develop a novel approach based on decision-theoretic active learning. On the important task of biomedical citation screening for systematic reviews, results on real data show that our method achieves consistent improvements over baseline strategies. To foster further research by others, we have made our data available online.

## Introduction

We investigate *finite-pool* data categorization (Wallace et al. 2010a), in which the objective is to exhaustively and accurately categorize a set of examples while minimizing cost. These categorizations will be performed either manually or using a classifier induced over the annotated examples. Training data thus serve a dual purpose: acquired labeled instances will be used not only to induce an accurate classification model, but also to reliably annotate some portion of the data. Because there is no separate training data, labeling errors are costly not only because they hamper performance of the learned classifier, but also because they are, by definition, misclassified items.

The defining characteristic of the finite-pool scenario is the focus on using a hybrid system – here composed of crowd workers, experts and a classification model – to label a specific, fixed set of instances. Thus this while a classification model is trained to reduce labeling effort, this learned model is not the primary output of interest. This scenario thus differs from much of the previous work in active learning, which has sought primarily to achieve strong classifier generalizability at low cost; in that case, misclassified training instances incur cost only insofar as they exacerbate classifier errors on unseen examples.

Crowdsourcing (Lease 2011; Lease and Alonso 2014) is an increasingly popular approach to acquiring annotations at low cost, often used to subsequently train classifiers. Despite demonstrated utility in this respect (Snow et al. 2008), the quality of annotations acquired from the crowd remains a concern. This is particularly true in domains in which label quality is paramount. In practice, this often results in a trade-off between cost and quality: one aims to make the best possible use of pricey domain experts and to efficiently capitalize on low-cost crowd annotations.

*Active learning* (Settles 2012), in which unlabeled items are selected for labeling cleverly (rather than at random) in an iterative and interactive process, is a natural fit for this scenario. By intelligently selecting items to label, active learning can economize annotator effort and realize greater predictive accuracy at lower cost.

In this paper, we aim to select both the item to be labeled and the expert to label it at each step in the learning process. In particular, we extend prior work to develop a decision theoretic active learning framework that jointly considers querying the crowd and domain experts for labels on examples deemed likely to be informative. Our contributions in this work can be summarized follows:

- As far as we are aware, this is the first empirical exploration of active learning with labels from both crowd workers and domain experts that uses real data (i.e., data collected for a real task from both types of labelers).
- We present and evaluate a new decision theoretic approach for this task that selects the labeler type (crowd worker or domain expert) to explicitly minimize the expected loss. Our approach is general, but here designed for scenarios in which one class is rare and misclassification costs are asymmetric (i.e., False Negatives are expensive).

The remainder of this paper is structured as follows. In the next section, we discuss our motivating scenario: citation[1] screening for biomedical systematic reviews. We then discuss related work to place our contributions in context. This is followed by presentation of our method and the baseline approaches to which we compare this to. We present our empirical setup and results in the subsequent section, and end with a discussion of future work.



[1] *Citation* here refers to paper abstracts and associated meta-data such as titles, author information and keywords.

## Motivating Application: Biomedical Systematic Reviews

*Evidence-based medicine* (EBM) aims to inform patient care using the entirety of the available evidence. The cornerstone of EBM is the *systematic review*, in which reviewers conduct a comprehensive synthesis of the evidence relevant to a precise clinical question, often via statistical meta-analysis of the outcomes (Barza, Trikalinos, and Lau 2009). Systematic reviews increasingly inform all levels of healthcare, from bedside practice to national policy.

Unfortunately, systematic reviews are extremely laborious (and hence expensive) to conduct. This problem has been exacerbated by the exponentially increasing biomedical evidence based; researchers can no longer keep pace with the literature (Bastian, Glasziou, and Chalmers 2010). Conducting a review involves following a specific sequence of steps that constitute a pipeline (Wallace et al. 2013b): (1) formulating a precise clinical question to be addressed; (2) designing a broad search strategy to retrieve all potentially relevant citations; (3) *screening* these with respect to their eligibility for inclusion in the review; (4) extracting the information of interest from the relevant (screened in) articles; and finally, (5) statistically synthesizing this information.

Our focus in this work is on step (3) of screening potentially relevant citations retrieved via a broad query. This usually involves experts (individuals trained in evidence-based medicine; usually MDs) reading the entire set of citations retrieved via database search to identify the small subset of these eligible for the review. Typically, this entails experts reading thousands of citations to identify the 5% or so deemed likely to be eligible for inclusion in the review.

This is a well defined problem where expert labels can be reasonably assumed to be correct. In practice, very high inter-annotator agreement of 96%-97% has been observed (Mateen et al. 2013), and "mistakes" tend to be one-sided in the form of False Positives.

Previous work has considered semi-automating this process via machine learning (Wallace et al. 2010b; Cohen et al. 2006). Furthermore, active learning – in which citations are selectively chosen for screening by an expert with the aim of inducing a better predictive model with less effort – has been shown to improve performance for this task (Wallace et al. 2010a). More recently, researchers have investigated crowdsourcing citation screening for systematic reviews via Mechanical Turk (Mortensen et al. 2015). In general, despite the task requiring some degree of biomedical knowledge, it was found that crowd workers were capable of providing screening decisions that correlated reasonably well with expert judgments at low cost. Nonetheless, these decisions were naturally not as high-quality as those made by (expensive) trained systematic reviewers.

Here we combine these lines of work by developing a single model that intelligently queries for supervision from domain experts and crowd workers to use as training data. The hope is that by so doing, one may minimize loss at lower cost than relying on either experts or crowd workers exclusively (or heuristically alternating between them).

There are two properties that make this aim difficult: asymmetric misclassification costs and high cost ratios (between crowd workers and experts). The former arises because in citation screening for systematic reviews it is crucial not to overlook relevant articles (i.e., citations that meet the inclusion criteria), because the entire purpose of conducting a systematic review is to be comprehensive. The latter property (high cost ratios) is due to the specialization necessary to provide highly accurate screening decisions. We believe these two properties are common to many real-world learning tasks, and thus that the methods we propose here generalize beyond the specific application of systematic reviews.

## Related Work

Active learning is now a well-studied sub-problem (Settles 2012). In general, active learning involves selecting training data intelligently, rather than at random, to maximize classifier performance on future examples. Many strategies have been proposed to realize this general aim. Perhaps the most popular of these is *uncertainty sampling* (Lewis and Gale 1994), in which the expert is asked to label the (as yet unlabeled) example about whose label the current model is least certain. Uncertainty can be quantified in a number of ways, depending on the underlying classification model being used. For example, if the classifier is a linear model, then the distance to the separating decision boundary provides a measure of certainty: one may simply select the item closest to the current decision boundary to be labeled.

Our work is directly based on previous work on "optimal" active learning, where the learner selects for labeling the example that minimizes the expected loss on the test set. Cohn, Ghahramani, and Jordan (1996) provided analytic solutions (and thus sampling strategies) to some simple problems of this general form. Roy and McCallum (2001) applied this idea to text classification and took a sampling (Monte Carlo) approach to make estimation tractable. They showed its improvement over classic uncertainty sampling and other baselines. Kapoor, Horvitz, and Basu (2007) later extended this method to incorporate variable cost labels.

In these "optimal" approaches, the core approach is to: (1) consider each item; (2) estimate its label; (3) add that label to the training set, (4) retrain the model and estimate the future error (subsequent to collecting the new label). This process is repeated for every unlabeled item, and the estimated future error is weighted by the estimated probability of that item taking the respective labels comprising the set of classes under consideration (e.g., 0 or 1 in a binary setting). This is naturally cast in a decision theoretic light: the active learner may be viewed as an agent reasoning about what actions it may take (which instances to request labels for) by considering the likely states it will be in after each possible action (the loss; typically predictive performance), and the expected costs of different actions (costs per label). This provides a principled, flexible framework to address the cost-quality trade-off inherent to active learning.

The strategy of combining active learning with crowd labelers has also been previously investigated, with encouraging results. For example, Laws, Scheible, and Schütze (2011) applied this strategy to two natural language processing tasks: named entity recognition and sentiment detection.

Elsewhere, Mozafari et al. (2014) proposed a general strategy for active learning using the (non-parametric) bootstrap.

Yan et al. (2011) investigated active learning from multiple labelers of varying expertise but did not consider their varying cost. Donmez and Carbonell (2008) introduced *proactive* learning, a generalization of active learning that relaxes unrealistic assumptions. Wallace et al. (2011) studied active learning from multiple experts, some less experienced than others, finding that one may rely on novice worker meta-cognition to decide when to appeal to experts. Computer vision researchers have also recently explored combining crowd and expert labels (Patterson et al. 2013).

In contrast to the above works, our approach here is explicitly designed for scenarios in which one aims to use domain experts, crowd workers and a classifier in conjunction to label a finite pool of examples with maximum accuracy at minimal cost. This alters the loss function, which must explicitly account for the costs associated with mislabeled training examples acquired from the crowd (in addition to errors made by the classifier). We do this by attempting to make optimal decisions at each step in the learning process, with respect to the expected loss, as we next describe.

## Methods

Our proposed active learning approach uses a decision-theoretic model (described next) to decide which instances should be labeled and by whom.

### Decision Theoretic Active Learning

Decision theory is a general framework for making decisions under uncertainty, characterized by:

- A set of states $S$ with a start state and perhaps an end state.
- A set of actions $A$.
- A transition function $T(s, a, s')$ that gives the probabilities of transiting to states $s'$ when performing action $a$ at state $s$ for all $s, s' \in S$ and $a \in A$.
- A loss function $L(s)$ for each state to be minimized.

Given this framework, one can opt to take the actions that minimize the expected loss, where this expectation marginalizes over the losses of all possible states, weighted by their respective probabilities (that is, the probabilities of transitioning to them after performing the action under consideration). Unfortunately, evaluating projected future actions (and associated expected losses) requires enumerating an exponentially large space of possible states, which is computationally intractable. We therefore use a common heuristic first-order approximation: we make a greedy decision that optimizes only the for expected loss in the immediate next state.

Recall that our aim is to maximize the quality of the labels (which translates to our loss) given some specified labeling budget. We define the state $s$ as: (1) the set of unlabeled items; (2) the set of labels we have collected so far (from both crowd workers and experts); and (3) the remaining budget. In the start state $s_0$, we assume all items are unlabeled except for an initial seed set of 100 crowd-labeled items. The

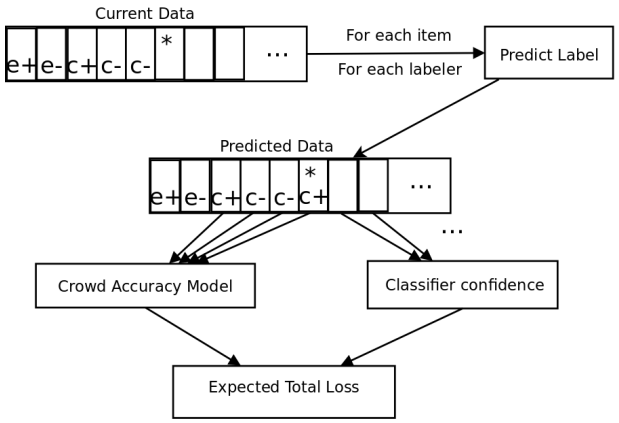



Figure 1: Illustration of our approach: 'e' and 'c' denote items with expert and crowd labels, where '+' and '-' mean positive and negative.* denotes the current item we consider.

end state is reached when the specified budget is exhausted. With regard to the set $A$ of actions, we consider two options:

1. Pick an unlabeled item and ask the crowd to label it (i.e., collect $k$ crowds labels for a cost of $k$ units).
2. Pick a crowd-labeled item and ask the expert to label it for a cost of $E$ units.

We do not consider bypassing the crowd (i.e., asking the expert to label an unlabeled item) because prior work (Wallace et al. 2011) has showed that always querying the lower cost labeler first (and only subsequently deferring to the expensive expert) works well in practice. We also simplify the action space and reasoning by requesting all 5 labels from the crowd at once. Given the price disparity between experts and crowd workers, the additional cost of acquiring all $k$ crowd labels (rather than $< k$) is marginal.

After performing an action, the state that we transition to depends on the response from the labeler, which we do not know beforehand. But we can use the predicted distributions over potential responses, thus providing the transition function $T$. The loss function here is a measure of the quality of the labels that we acquire for all items (including those labeled by a classifier trained on using the labels acquired so far). For example, a simple loss function would just be the number of misclassifications. In general, when we request a label from the crowd, we collect multiple annotations to mitigate noise. In this work, we simply pool these responses using majority voting; we use this straight-forward aggregation strategy because it has been shown to be competitive despite its simplicity (Sheshadri and Lease 2013).

The loss is a function of misclassification counts (False Positives and False Negatives), but in practice tallying these requires gold standard reference labels that we do not have during active learning. Therefore, in place of this reference set of labels, we use the predicted 'true' labels for each item when calculating the expected loss (factoring in associated uncertainty around these predictions). We note that care must be taken here in calibrating probability estimates

from a sample collected via active learning: by definition this will not constitute an independent and identically distributed (i.i.d.) sample of the item space. We address this via inverse-weighting, as described further below.

The challenges inherent to our approach can be decomposed into three sub-tasks:

1. Predict the response to an item from the crowd or expert.
2. Predict the true label of unlabeled items.
3. Predict the true label of items with only crowd labels (we refer to this model as the crowd accuracy model).

## Proposed Active Learning Approach

We use probabilistic classifiers for these tasks. Concretely, these are trained on the (crowd and expert) labeled portion of the dataset and then applied to unlabeled items, providing an estimated distribution over the labels of those items which we use a proxy for the first two task above. For the crowd accuracy model, we need to take into account crowd fallibility. We consider a simple model where crowd decisions on a given item are assumed i.i.d., conditional on the true label. That is, we assume that each crowd worker gives a noisy version of the true label and that the accuracy of this response is contingent on what the true label is. Thus for binary labels, the conditional distributions are modeled as Bernoulli random variables specifying the probability of responding with 0 or 1. These class-conditional parameters can be estimated from the items for which we have both crowd and expert labels, where the latter is taken as ground truth. Furthermore, we can incorporate an informative prior to reflect *a priori* beliefs regarding crowd worker accuracies.

**Algorithm 1**: Estimate the expected loss of a state

**INPUT**:

- $D_E$ = items with expert labels.
- $D_C$ = items with only crowd labels.
- $D_U$ = items with no labels.
- Function $L(\hat{y}, y)$ = the loss of returning label $\hat{y}$ for an item of true label $y$.
- Crowd Accuracy Model $P_C(x, K, y)$ = probability that the true label for item $x$ is $y$, given the set of crowd labels $K$.
- $Y$ = set of possible labels, $\{0, 1\}$ for binary data.

1: $\text{Clf} \leftarrow \textsc{Train}(D_E, D_C)$ ▷ Train classifier
2: $\text{Loss}_C \leftarrow 0$ ▷ Loss for crowd-labeled examples
3: **for** $(x, K) \in D_C$ **do**
4: $\quad \hat{y} \leftarrow \textsc{ConsensusLabel}(K)$ ▷ Majority vote
5: $\quad \text{Loss}_C \leftarrow \text{Loss}_C + \sum_{y \in Y} P_C(x, K, y) L(\hat{y}, y)$
6: $P_U \leftarrow \textsc{PredictProb}(\text{Clf}, D_U)$
7: $\text{Loss}_U \leftarrow 0$ ▷ Loss for unlabeled examples
8: **for** $x \in D_U$ **do**
9: $\quad \hat{y} \leftarrow \textsc{PredictLabel}(\text{Clf}, x)$
10: $\quad \text{Loss}_U \leftarrow \text{Loss}_U + \sum_{y \in Y} P_U(x, y) L(\hat{y}, y)$
11: **return** $\text{Loss}_C + \text{Loss}_U$

Together, the classifier and crowd accuracy models are sufficient to estimate the loss expected in an arbitrary state. **Algorithm 1** provides more concretely the steps we take to estimate this. Note that the expected loss incurred for items labeled only by crowd members and those that are as-yet unlabeled are computed separately. We assume expert infallibility (i.e., zero loss for expert labels) as a simplifying but often nearly true assumption, evidenced by the extremely high inter-annotator agreement rates cited earlier (Mateen et al. 2013). In line 6, the (class conditional) crowd accuracy model is used to predict the probability that each crowd labeled item is correct. This estimate is in turn used to weight the corresponding contribution to the loss. In Line 9, the classifier trained on labeled items is used to predict class membership probabilities for unlabeled items. Intuitively, the expected loss for a given item is smaller when the crowd accuracy model (when we have crowd labels) or the classifier (for unlabeled items) agrees with the label in question.

Given the algorithm just described to estimate the loss at an arbitrary state, we can now evaluate the expected loss for each potential action. The active learning strategy naturally follows from this: the selected action will be a function of the estimated losses incurred by taking each action, scaled by the associated labeling cost. This strategy explicitly aims to maximize loss reduction per cost unit.

**Algorithm 2**: Decision-making (active learning)

**INPUT**: Same as Algorithm 1

1: $\text{Clf} \leftarrow \textsc{Train}(D_E, D_C)$
2: $P_U \leftarrow \textsc{PredictProb}(\text{Clf}, D_U)$
3: **for** $x \in D_U$ **do** ▷ Consider querying the crowd
4: $\quad L[C, x] \leftarrow 0$
5: $\quad$ **for** $y \in Y$ **do**
6: $\quad\quad \text{ExpLoss} \leftarrow \text{Algorithm1}(D_U - x, D_C + (x, y))$
7: $\quad\quad L[C, x] \leftarrow L[C, x] + P_U(x, y)\text{ExpLoss}$
8: **for** $(x, K) \in D_C$ **do** ▷ Consider querying the expert
9: $\quad L[E, x] \leftarrow 0$
10: $\quad$ **for** $y \in Y$ **do**
11: $\quad\quad \text{ExpLoss} \leftarrow \text{Algorithm1}(D_C - x, D_E + (x, y))$
12: $\quad\quad L[E, x] \leftarrow L[E, x] + P_C(x, K, y)\text{ExpLoss}$
13: $\text{CL} \leftarrow \text{Algorithm1}$ ▷ Current Loss
14: $\text{Score} \leftarrow []$ ▷ Hash labeler/item pairs to scores
15: $\text{Score}(C, x) \leftarrow (\text{CL} - L[C, x])/\text{Cost(C)}) \forall x \in D_U$
16: $\text{Score}(E, x) \leftarrow (\text{CL} - L[E, x])/\text{Cost(E)}) \forall x \in D_C$
17: **return** Score

**Algorithm 2** presents the core of this decision making procedure, which calculates scores (expected losses over costs) for each state using Algorithm 1 as a subroutine. We consider taking the possible action for each item not yet labeled by the expert. Thus each item in the two candidate pools ($D_E$ and $D_U$) is considered, and we simulate adding these to the corresponding labeled sets with each (hypothetical) label in turn. We then estimate the loss that would be realized if this hypothetical label were accurate by Algorithm 1 and weight the contributions of these expected losses with the estimated probabilities that the hypothetical label is in-

deed the 'true' label (using the crowd accuracy model or the classifier). For a given item/action pair, the expected loss is then the average of the losses estimated for each possible label, weighted by the estimated probabilities of the said labels. We denote these by $L(C, x)$ and $L(E, x)$, denoting expected losses associated with querying the crowd and expert for item $x$, respectively. Finally, the score for each of these item/action pairs is is simply the expected reduction in loss divided by the cost of that action.

## Accounting for Active Sampling Bias

Given the preceding discussion, the optimal decision at any given point is to simply select the action with the highest score. However, rather than taking this deterministic approach, we instead opt to select an action with probability proportional to its score. That is, the scores calculated in Algorithm 1 are normalized to a probability distribution over actions and we select a specific action from this distribution. The motivation for this stochastic approach is two-fold. First, it enables continuous exploration of the space, which may avoid myopic focus on a single area of the item space. Perhaps more importantly, it lets us perform bias correction for items selected by active learning, which we now discuss.

Training sets collected via active learning exhibit sampling bias by definition: an i.i.d. sample is just standard (passive) learning. In practice, if one is performing uncertainty sampling, then items selected during active learning will tend to be closer to the current decision boundary. The distribution of such items may be considerably different from the population of all items. For example, consider the case of binary imbalanced datasets, which contain far fewer items from one class (the minority class) than another (the majority class). Because uncertainty sampling tends to select items close to the decision boundary, it follows that minority examples will be over-represented in the resultant dataset.

Unfortunately, most learning algorithms assume that the training set is an i.i.d. sample of the population. The bias induced by active learning can severely affect classifier performance. Our solution to this follows previous efforts to address this issue in active learning and specifically is based on Importance Weighted Active Learning (Beygelzimer, Dasgupta, and Langford 2009). The idea, known generally as Horvitz-Thompson estimation (1952), is to weight items by the inverse of the probabilities of having been selected:

$$w_i = \frac{1}{p_i} \propto \frac{1}{\text{score}(R, x_i)} \tag{1}$$

where $R \in \{C, E\}$, depending on whether $x_i$ was labeled by the crowd or the expert (see Algorithm 2). The weights $w$ can then be used to correct an estimator. Continuing our example of the imbalanced dataset, suppose we would like an estimate of the proportion of positive items in the population of items, derived using the items selected via uncertainty sampling. A standard maximum likelihood estimator (over an i.i.d. sample) would simply calculate the proportion of positive items selected. However, as noted above and in previous work (Wallace et al. 2013a), this will almost certainly be inflated in the actively collected sample. We can correct this using the following weighted estimator:

$$\hat{\text{p}} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

where $y_i$ is the binary label for item $i$. This estimator gives higher weights for items with lower probability of having been selected and therefore intuitively tends back to the estimate that might be taken over an i.i.d. sample.

In our case, we use item-weighting during model parameter estimation, where weights are set to $w_i$ (defined above). Intuitively, the classifier will be biased to finding parameters that correctly classify those items with larger weights, because it incurs greater penalty during training when these are misclassified. Our aim here is to improve the probability estimates that we rely on to calculate the expected loss; if these are unreliable due to sampling bias, then it follows that the expected loss estimates will also be unreliable and hence the entire decision theoretic approach may fail.

## Estimating the Accuracy of Crowd Workers

The last component in our approach is the Crowd Accuracy Model. For concreteness and clarity, we present our model for the case of binary labels, but the approach may be trivially generalized to multiclass problems. We make the assumption that, conditioned on the true (reference) label, crowd members independently and identically give the correct answer with some (latent) probability to be estimated. Put another way, the answers of $k$ crowd members, given the true label $\in \{0, 1\}$, is a Binomial random variable $\text{Bin}(k, p)$ where $p$ is the parameter of interest. We have one such variable for each possible (true) label. Estimation of $p$ for both label cases is straightforward: we take a smoothed maximum likelihood estimate using the set of items labeled both by crowd workers and the expert. Specifically, let $y_i^c$ be the crowd label and $y_i^*$ be the true label. We have:

$$P(y_i^c = y' | y_i^* = y) = \frac{c(y', y) + \alpha_{y',y}}{\sum_{\tilde{y} \in \{0,1\}} [c(\tilde{y}, y) + \alpha_{\tilde{y},y}]} \tag{3}$$

where $c(y', y)$ is the observed number of items that the expert labeled as $y$ and a crowd member labeled as $y'$. The $\alpha$s are smoothing parameters (which may be viewed as priors on counts, assuming a Bayesian view). Using this model, we can predict the correct label of items from $k$ crowd labels $\{y'_1 ... y'_k\}$ as follows:

$$P(y_i^* = y | y_i^{c1} = y'_1 ... y_i^{ck} = y'_k) = \tag{4}$$

$$\frac{\prod_{j=1}^{k} P(y_i^{cj} = y'_j | y_i^* = y) P(y_i^* = y)}{P(y_i^{c1} = y'_1 ... y_i^{ck} = y'_k)} \tag{5}$$

**Figure 1** schematically depicts our overall approach. In practice, this approach is computationally intractable because it requires considering each possible action for each item by calculating the expected loss in each possible new

state, which entails retraining the classifier to estimate probabilities to evaluate the loss. Fortunately, various approximations can be used to make the approach practically feasible, which we discuss in detail below.

## Implementation Details

For the base probabilistic classifier, We adopted regularized Logistic Regression as implemented in the Scikit-Learn library (Pedregosa et al. 2011) using Stochastic Gradient Descent (Bousquet and Bottou 2008).

Given features $\{f_k^{(i)}|k = 1..F\}$ for each item $i$, the classifier defines the conditional probability of the label $Y^{(i)}$ as:

$$Pr(Y^{(i)} = 1|f^{(i)}) = \text{sigm}\left(\sum_{k=1}^{F} w_k f_k^{(i)}\right) \quad (6)$$

Given a labeled training set, Gradient Descent minimizes its negative conditional log likelihood, with regularization:

$$\mathcal{L}(w) = -\sum_{i=1}^{n} \log(Pr(Y^{(i)}|f^{(i)})) + \alpha R(w) \quad (7)$$

by moving the weights $w$ a small step in the direction of steepest descent, which is the negative gradient. The empirical loss is traded off against a regularization penalty $R$ that attempts to keep weights small. The degree of regularization is controlled by the parameter $\alpha$. Here we used a default $\alpha$ value of $0.0001$ and $L_1$ regularization, defined as:

$$R(w) = ||w||_1 = \sum_{k=1}^{F} |w_k| \quad (8)$$

A feature of Logistic Regression trained by Stochastic Gradient Descent that is important for our approach is its support for incremental training, i.e., efficiently updating the weights after new examples become available. This is critical for the performance since our approach involves retraining the classifier after new item(s) with labels are added.

Although we aim for a principled approach, some aspects of this problem necessitate heuristics to address practical problems. Heuristics were developed using only one of the four datasets and were not tuned to optimize performance.

Firstly, there is the aforementioned problem of computational intractability from having to enumerate all possible next states. We follow (Roy and McCallum 2001) and adopt a pruning strategy to reduce the search space, considering only the 100 items for each action type (query crowd and query expert) about which the model is least certain. Secondly, for the querying expert action, we approximate the expected loss reduction of the action by the expected loss of the crowd-labeled item. In practice, the main improvement to be realized through querying the expert is to increase certainty about the crowd labeled item. These heuristics yielded a very practical run-time, typically requiring only a few seconds to make each decision on a 3.40GHz machine.

Another practical problem we encountered is that the predicted loss reduction of querying the crowd for an unlabeled item can be very noisy. Indeed, **Figure 2** illustrates

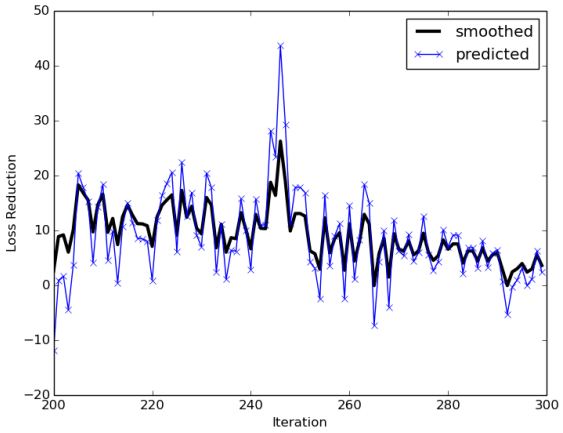



Figure 2: An example of the maximum predicted and smoothed loss reduction (over all unlabeled examples) of querying the crowd. The values plotted are proportional the maximum score of querying the crowd for any item.

this point: one can see many (positive and negative) spikes in the plot. Our model predicts that querying the crowd for a given unlabeled item can greatly reduce the loss, but at some point the action slightly increases the loss. Recall that the action of querying the crowd for an unlabeled item incurs a small expected loss of the crowd being incorrect, and this is effectively traded off against the expected gain of receiving a correct label from the crowd on the given item. We believe the latter estimate is noisy, in part, because we are only looking one step ahead. To mitigate this problem, we smooth the estimate through interpolation, thus incorporating observed reductions in loss resulting from querying the crowd for labels. Specifically, we take the final loss reduction estimate as:

$$0.5\text{LR} + 0.5\frac{1}{T}\sum_{i=1}^{T} P_{T-i}\lambda^i \quad (9)$$

where LR is the current predicted reduction, $T$ is the number of iterations taken, $P_{T-i}$ is the reduction in loss $i$ iterations in the past, and $\lambda = 0.99$ gives more weight to the more recent reduction. This technique is applied separately for both actions of querying either the crowd or an expert.

# Experiments

## Datasets

We use four systematic review datasets for our experiments. All of these comprise the screening decisions made by experts (professional systematic reviewers) regarding the relevance of citations to the corresponding review. These are binary judgements, because published studies either will or will not meet the specified inclusion criteria for a given review. Characteristics of these datasets are presented in **Table 1**. For each of these datasets, we collected crowd labels via Amazon's Mechanical Turk platform for questions which

| Dataset | Number of citations | Deemed relevant (%) |
|---|---|---|
| Proton Beam | 4,749 | 243 (5.1%) |
| Appendicitis | 1,664 | 242 (14.5%) |
| DST | 8,071 | 183 (2.3%) |
| Omega 3 | 5,774 | 310 (5.3%) |

Table 1: Characteristics of the systematic review datasets used. We report the total number of citations screened for each review and the number deemed relevant (i.e., $y = 1$).

taken together constitute the inclusion criteria (Mortensen et al. 2015). For every abstract, we collected 5 such crowd decisions. This dataset is freely available online[2].

We refined our approach using two of the four datasets, `Proton Beam` and `Appendicitis`, as *development* data. We held out the other two datasets, `DST` and `Omega 3`, for final testing; this helped safeguard the generality of our findings by preventing us from over-engineering an approach for the specific datasets used.

## Experimental Setup

For each citation, we derived unigram TF-IDF features from the corresponding title, abstract and keywords. We did not remove stop-words. For each citation, we obtained a 'gold label' from the expert and 5 labels from the crowd. We simulated active learning as follows.

The algorithm being tested is given access to all feature vectors for all items, and a small seed set of all (i.e., 5) crowd labels for 100 randomly selected items. The learner is allotted a budget to be spent on requesting additional labels. At each step: (1) the learner selects an item and labeler (crowd or expert); (2) the corresponding label is exposed to the learner; and (3) its budget is updated and the underlying model is retrained. This process proceeds until the budget has been exhausted.

Naturally, the cost of querying an expert is (considerably) more expensive than acquiring labels from the crowd. We set the cost of an individual crowd label to 1 unit and the cost of an expert label to $E$ units, so querying the expert is $E$ times as expensive as querying a single crowd worker.

Evaluation is performed at each step in the simulated active learning process. Specifically, we measure: (1) the quality of the acquired labels and the classifier induced on this training set, and (2) the cost incurred so far on these labels. After a new item is selected and the label revealed, the classifier is retrained and used to predict labels (and associated probabilities) for the as-yet unlabeled examples. For each item with only crowd labels, a consensus label is inferred by majority vote. Next, all of the collected and predicted labels are compared to the true labels. The loss is then calculated as a weighted sum of False Positives and False Negatives:

$$\text{Loss} = \text{FP} + R \times \text{FN} \tag{10}$$

The loss ratio $R$ indicates that missing one relevant document (i.e., a False Negative) is as costly as including $R$ irrelevant documents (i.e., False Positives). In the systematic review domain, these rare positive examples are the relevant citations, and overlooking them is extremely expensive because it may undermine the comprehensiveness of the systematic review being conducted.

In our main experiments, we set cost ratio $E = 100$ and loss ratio $R = 10$ to realistically reflect our motivating scenario (outlined above). The cost ratio $E = 100$ captures the observation that a typical (Mechanical Turk) crowd worker might earn $\approx$ \$1.5/hour, while a trained physician might earn $\approx$ \$150/hour.

We compare 4 algorithms in our experiments:

- **US-Crowd**. Uncertainty Sampling: crowd labels only
- **US-Expert**. Uncertainty Sampling: expert labels only
- **US-Crowd+Expert**. Use Uncertainty Sampling to select an item for the crowd to label. If crowd members are not unanimous, then send that item to the expert to relabel.
- **Decision Theory.** Use our approach, as described above.

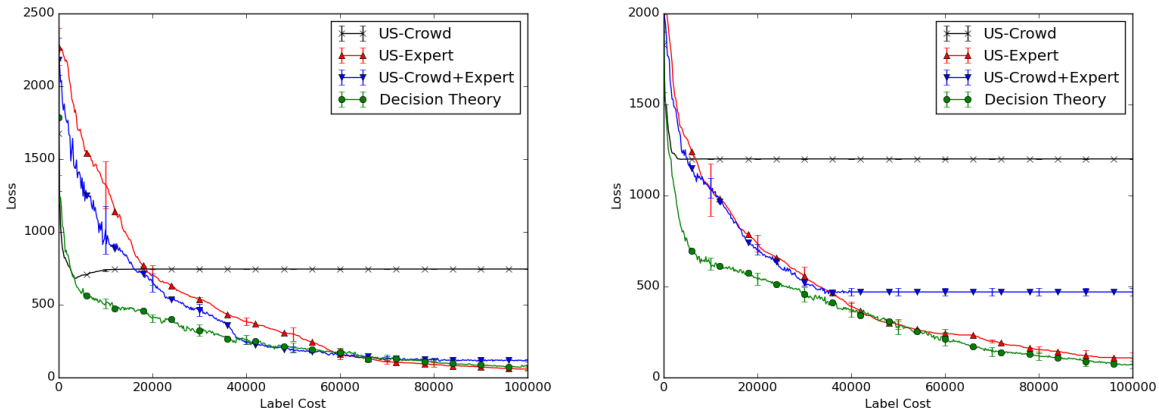
The uncertainty is measured using the classifier's predicted probability. That is, we select examples with predicted probability closest to 0.5.

**Figure 3** presents results for `Proton Beam` and `Appendicitis`. Results summarize 5 independent runs in which all strategies had access to the same randomly selected (for each run) set of seed crowd-labeled examples. In both datasets, one can see substantial improvement realized by the proposed decision theoretic approach, compared to other algorithms. US-Crowd, which uses only crowd labels, does manage to reduce the loss rapidly in the beginning, but then stops improving later on. At the outset, when exploration is most valuable, spending a small amount of money to collect many crowd labels is quite beneficial. At some point, errors in these noisy labels begin to increase loss.

US-Expert performs well in the end, but it is not efficient when the budget is limited. If we allow higher loss then other algorithms can considerably reduce labeling cost. US-Crowd+Expert makes use of crowd labels and also defers to the expert when necessary. It is conceptually simple but has limitations. As can be seen, toward the beginning, asking the expert for every item in which the crowd is not unanimous is not efficient. And when many labels have been collected (near the end of the learning process), there remain items about which the crowd is unanimously wrong (evidencing systematic bias between crowd vs. expert populations). This can be seen clearly in the Appendicitis dataset.

One can imagine the ideal strategy of acquiring crowd labels at the start of the learning process and then making the gradual transition to deferring to the expert at some later point. Our approach captures this intuition in a principled fashion. In the beginning, when we have very few labels, acquiring many labels provides the classifier with new training examples, which helps the model make better predictions on the large set of as-yet unlabeled items. The resulting expected loss reduction outweighs the small expected loss associated with the uncertainty of the crowd labels. As we collect more labels, the classifier becomes increasingly confident about the unlabeled items and the benefit of acquiring more crowd labels diminishes. Expert labels become more valuable at this point, justifying their higher cost. The point

[2] github.com/bwallace/crowd-sourced-ebm




(a) Proton Beam

(b) Appendicitis

Figure 3: Cost v. loss curves of the four approaches on the Proton Beam and Appendicitis datasets. The result for each is averaged over 5 runs and error bars of one standard deviation are shown to illustrate variance. When the algorithm runs out of new labels, the lines are extended to the end of the X-axis. For example, the US-Crowd line for Proton Beam ends at approximately 25000.

at which this happens depends on many factors, including the difficulty of the screening task (some of these systematic reviews require less technical knowledge than others), the accuracy of the crowd, and the cost and loss ratios. All of these are explicitly modeled in our approach.

**Figure 4** presents the results on our two *held-out* datasets, `DST` and `Omega 3`. Our decision theoretic approach again substantially outperforms other approaches in the beginning. However, the results are more mixed on these two datasets compared to the above. On DST, US-Expert and US-Crowd+Expert catch up to Decision Theory after the 30000 cost mark. For a period, US-Expert slightly outperforms Decision Theory, until a bit after the 100000 cost unit mark, at which point Decision Theory performs comparably to US-Expert, and better than US-Crowd+Expert. In the case of Omega 3, after another strong showing at the start of the process, Decision Theory is outperformed by US-Crowd+Expert for a part of the curve but then ultimately achieves comparable performance. It is worth noting that US-Crowd+Expert performed quite poorly in the case of both Appendicitis and DST, so this approach is rather inconsistent. By contrast, our proposed approach consistently performs as well as other strategies and frequently outperforms all baseline approaches for at least some portion of the cost-loss curve.

In sum, results on the two *held-out* datasets suggest that (1) our Decision Theory approach is consistently one of the best strategies across datasets, and, (2) it is by far the most effective strategy when one has a small budget.

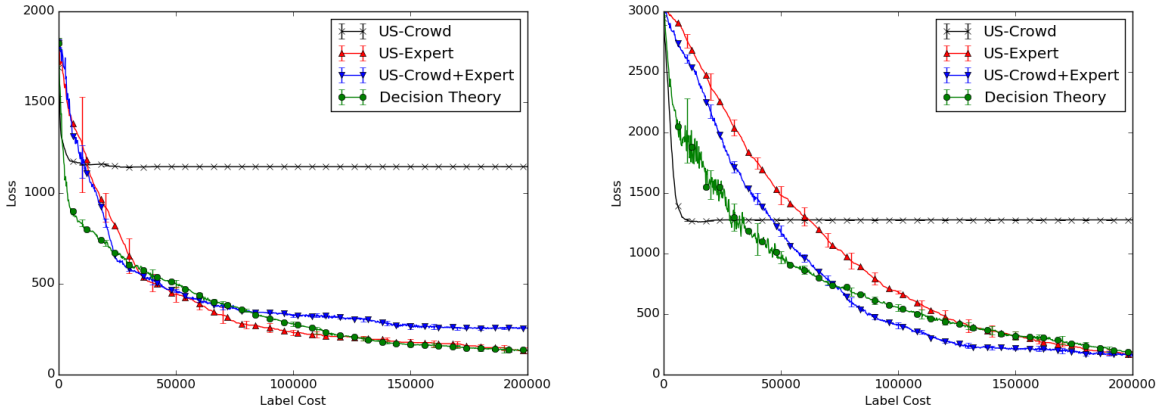
## Other Settings

We developed our approach with particular focus on our motivating application of biomedical citation screening for systematic reviews. This application is characterized by a relatively high cost of false negatives and also relatively expensive domain experts (compared to crowd worker costs). However, our approach is general and can be adapted to alternative scenarios by appropriately adjusting the loss and expert cost ratio parameters (i.e., setting $R$ and $E$ to reflect the trade-offs inherent to a target domain).

To demonstrate this, we again apply our approach to our two development datasets of Proton Beam and Appendicitis, but assuming different values of $E = 25$ (instead of 100) and $R = 1$ (instead of 10). In the case of the former, the proposed approach fares rather well, ultimately outperforming all other strategies when the budget is exhausted (although showing a dip in performance on the Proton Beam dataset around a cost of 10000). However, results when we vary the cost ratio (of false positives to false negatives) are more mixed. In these cases, our decision theoretic approach does, for both datasets, ultimately end up as one of the top two strategies. However, in both cases it is outperformed by other strategies earlier in the learning process.

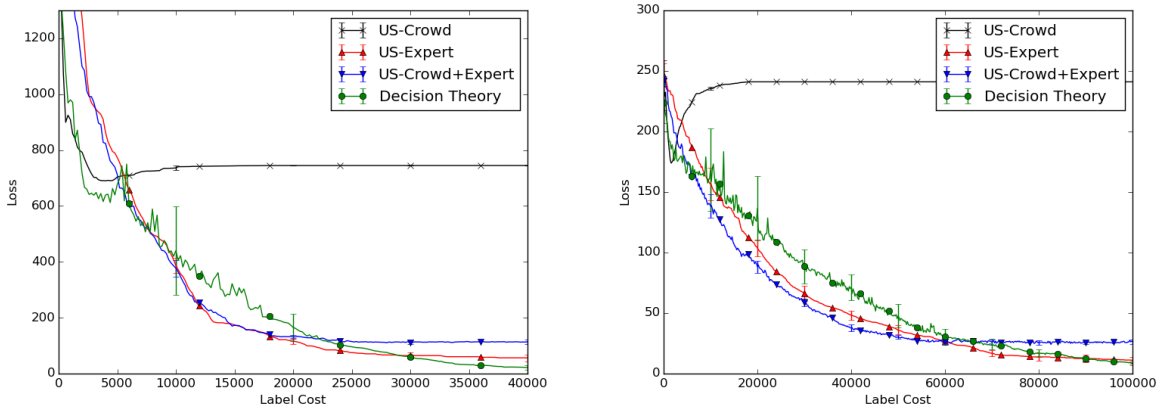
## Conclusions and Future Work

This is the first work that we are aware of to explicitly consider an approach for active learning from a domain expert and crowd simultaneously. We are interested in scenarios in which one aims to exhaustively label a finite pool of examples with a limited budget, using a combination of crowd workers, domain experts and machine learning. To explore this setting, we used real data – comprising labels from a domain expert and from crowd workers – for the task of biomedical citation screening. We have made this data available for further research. We have extended previous work in decision theoretic "optimal active learning" to allow us to formally reason about which item to next seek a label for and, jointly, whether to query crowd workers or a domain expert for said label. Results demonstrate that the proposed approach performs well for biomedical citation screening.

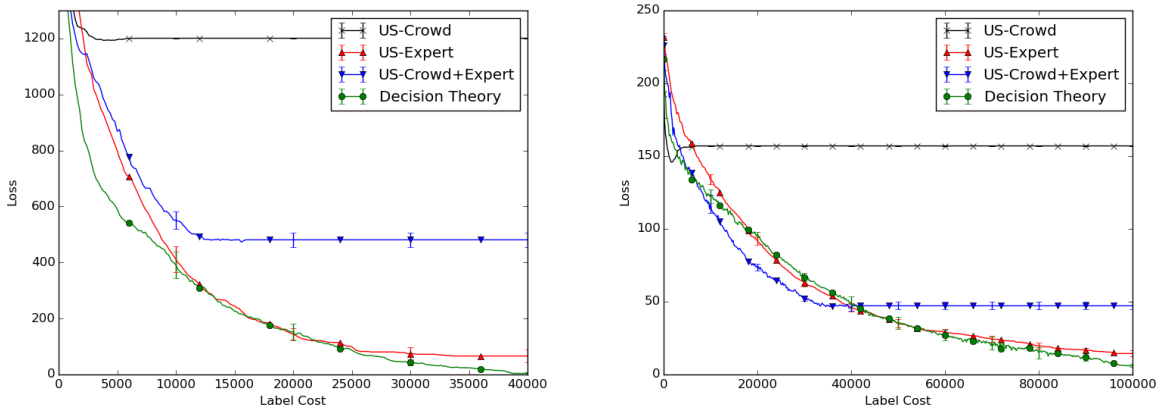



(a) DST (b) omega3

Figure 4: Results (cost v. loss curves) on two held out datasets, from the DST and Omega-3 reviews.




(a) Proton Beam (E=25, R=10) (b) Appendicitis (E=100, R=1)




(c) Appendicitis (E=25, R = 10) (d) Appendicitis (E=100, R=1)

Figure 5: Loss-cost curves for Proton Beam and Appendicitis datasets under different cost ratio $E$ and loss ratio $R$ settings.

Moreover, we have shown the approach can be generalized to different settings and maintain competitiveness.

In future work, we plan to better model individual crowd workers by considering worker identity and explicitly modeling noise in crowd labels. This could include routing labeling tasks to specific workers based on predicted performance (Jung and Lease 2015). Moreover, we will explore multi-step look-ahead to make better decisions (with care to avoid intractability). Another direction will be to use the model to provide label quality assurance or guarantees.

## Acknowledgments

We thank Ray Mooney and the anonymous reviewers for thoughtful comments and suggestions. We also thank the online crowd contributors who have made this study possible and enabled research on crowdsourcing and human computation to exist and flourish. This study was supported in part by National Science Foundation grant No. 1253413, DARPA Award N66001-12-1-4256, and IMLS grant RE-04-13-0042-13. Any opinions, findings, and conclusions or recommendations expressed by the authors are entirely their own and do not represent those of the sponsoring agencies.

## References

Barza, M.; Trikalinos, T. A.; and Lau, J. 2009. Statistical considerations in meta-analysis. *Infectious disease clinics of North America*.

Bastian, H.; Glasziou, P.; and Chalmers, I. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine* 7(9):e1000326.

Beygelzimer, A.; Dasgupta, S.; and Langford, J. 2009. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 49–56. ACM.

Bousquet, O., and Bottou, L. 2008. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, 161–168.

Cohen, A. M.; Hersh, W. R.; Peterson, K.; and Yen, P.-Y. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13(2):206–219.

Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1996. Active learning with statistical models. *Journal of artificial intelligence research*.

Donmez, P., and Carbonell, J. G. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*.

Horvitz, D. G., and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260):663–685.

Jung, H. J., and Lease, M. 2015. Modeling Temporal Crowd Work Quality with Limited Supervision. In *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*.

Kapoor, A.; Horvitz, E.; and Basu, S. 2007. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI*.

Laws, F.; Scheible, C.; and Schütze, H. 2011. Active learning with amazon mechanical turk. In *Proceedings of the conference on empirical methods in natural language processing*, 1546–1556. Association for Computational Linguistics.

Lease, M., and Alonso, O. 2014. Crowdsourcing and human computation, introduction. *Encyclopedia of Social Network Analysis and Mining (ESNAM)* 304–315.

Lease, M. 2011. On Quality Control and Machine Learning in Crowdsourcing. In *Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI*, 97–102.

Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 3–12. Springer-Verlag New York, Inc.

Mateen, F. J.; Oh, J.; Tergas, A. I.; Bhayani, N. H.; and Kamdar, B. B. 2013. s versus titles and abstracts for initial screening of articles for systematic reviews. *Clinical epidemiology* 5:89–95.

Mortensen, M. L.; Wallace, B. C.; Adam, G. P.; Trikalinos, T. A.; and Kraska, T. 2015. Crowdsourcing citation screening for systematic reviews. *In submission*.

Mozafari, B.; Sarkar, P.; Franklin, M.; Jordan, M.; and Madden, S. 2014. Scaling up crowd-sourcing to very large datasets: A case for active learning. *Proceedings of the VLDB Endowment* 8(2).

Patterson, G.; Horn, G. V.; Belongie, S.; Perona, P.; and Hays, J. 2013. Bootstrapping fine-grained classifiers: Active learning with a crowd in the loop. In *NIPS Workshop on Crowdsourcing*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Roy, N., and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *In Proc. 18th International Conf. on Machine Learning*.

Settles, B. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*.

Sheshadri, A., and Lease, M. 2013. Square: A benchmark for research on computing crowd consensus. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2010a. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 173–182. ACM.

Wallace, B. C.; Trikalinos, T. A.; Lau, J.; Brodley, C.; and Schmid, C. H. 2010b. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 11(1):55.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2011. Who should label what? instance allocation in multiple expert active learning. In *SDM*, 176–187. SIAM.

Wallace, B. C.; Dahabreh, I. J.; ; Moran, K. H.; Brodley, C. E.; and Trikalinos, T. A. 2013a. Active literature discovery for scoping evidence reviews: How many needles are there? In *KDD workshop on data mining for healthcare (KDD-DMH)*.

Wallace, B. C.; Dahabreh, I. J.; Schmid, C. H.; Lau, J.; and Trikalinos, T. A. 2013b. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. *Journal of comparative effectiveness research* 2(3):273–282.

Yan, Y.; Fung, G. M.; Rosales, R.; and Dy, J. G. 2011. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 1161–1168.