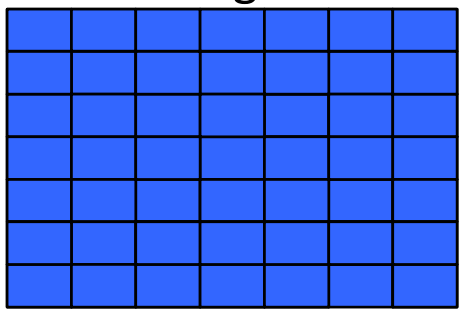
Leveraging Online Labor Markets for Machine Learning: Cost-Effective Labeling with Imperfect and Biased Humans

Maytal Saar-Tsechansky

Joint work with Ruijiang Gao, Tomer Geva, and Harel Lustinger

Data is fundamental to machine learning

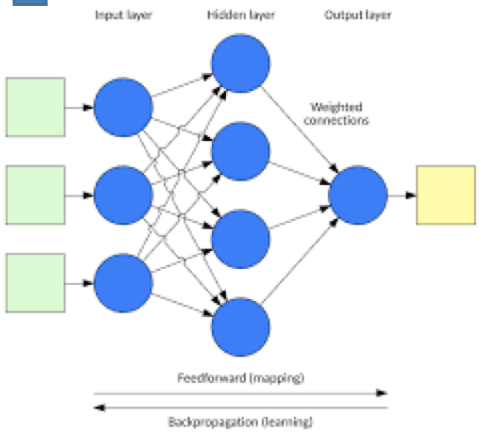
Training Data



A machine learner that also infers about what data to acquire for learning

Model induction

Predictive Model



Data is a fundamental ingredient for machine learning

Human Labeling for Machine Learning: A Bottleneck

Supervised machine learning requires **labeled** training instances (dependent variable value is known)

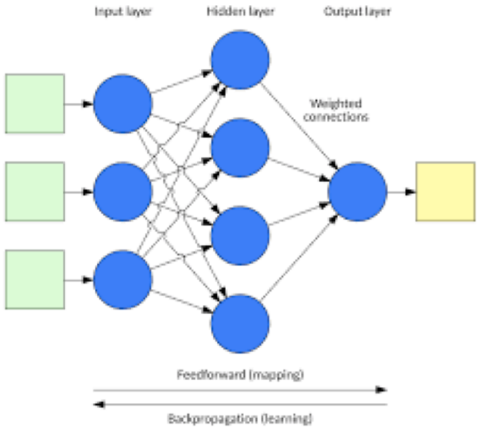
Human labeling: For many tasks (image, text prediction) **human intelligence is necessary for labeling training data** (e.g., what's the action in a video/image? Is a text humorous?).

Does the post contain hate speech?

	Indpendent Variables				Dependent Variable
ID	X1	X2	X3	…	Y
100	1	1	0		?
200	1	0	0		?
300	1	0	1		?
400	0	0	1		?
…	…	…	…	…	…

(Supervised) Machine Learning

Predictive Model



Online Labor Markets for Human Labeling: Great Opportunities & Challenges

The Opportunity: Alleviating human-labeling bottleneck

Online Labor Markets/Platforms (e.g., Amazon Mechanical Turk) offer **unprecedented immediacy and scope** for label acquisition

The Challenge:

- **Costs**: Cumulatively, the cost of human labeling is significant.
- **Imperfect Humans:** Human labeling is noisy

Our research:

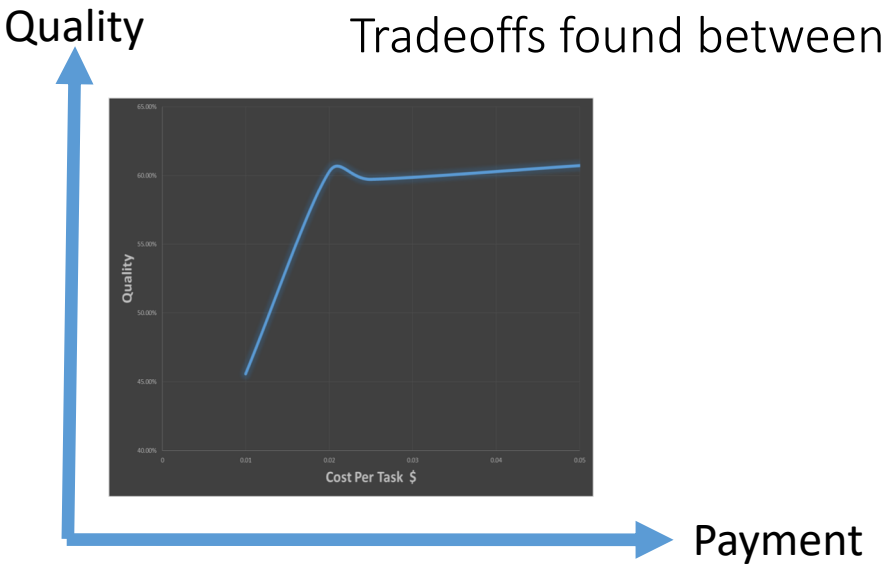
What to pay labelers so as to produce the best model given a budget?

Challenges: In markets, human labeling quality can vary with payment

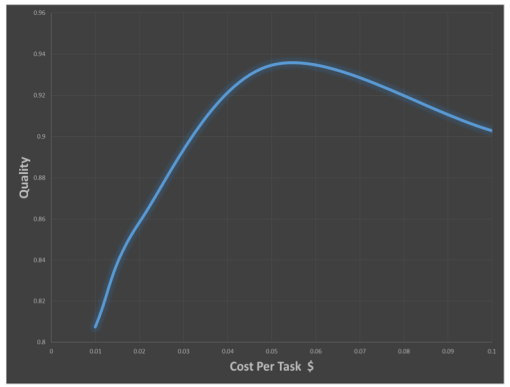
Field experiments with human labelers found:

- Different payments can lead to different labeling quality (and thereby model induction)
- Different tradeoffs between payment and quality arise in different contexts (e.g., different tasks, at different times, etc.)
- **No theory to predict what tradeoff will arise in any arbitrary setting**

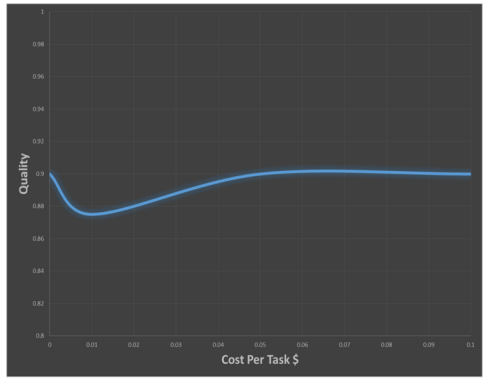
Tradeoffs found between payment and human labeling quality in field experiments:



"Asymptotic"
(Kazai, 2011; Kazai et al. 2013)



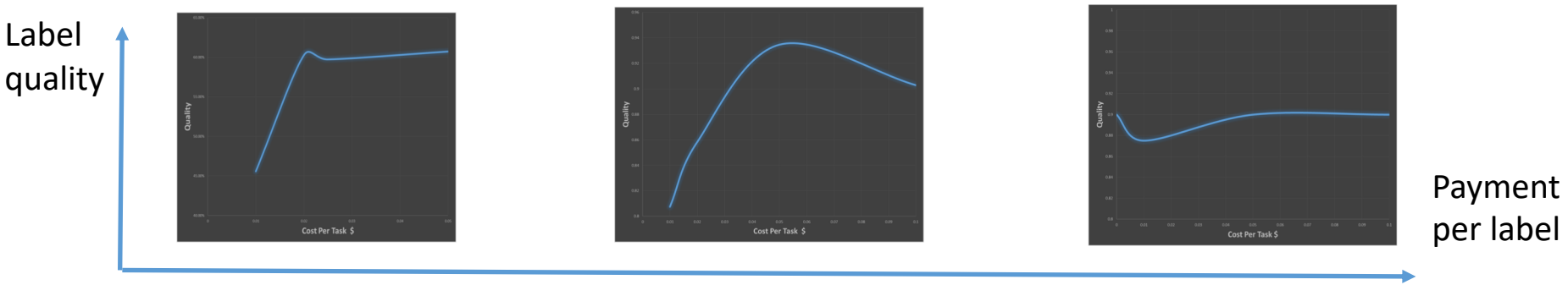
"Concave"
(Feng et al., 2011; Kazai, 2011; Kazai et al., 2013)



"Fixed"
(Mason and Watts, 2010)

Key properties of the problem

1. Unknown Cost/Quality tradeoff: A given payment yields unknow quality (and the same payment may produce different label quality over time)



2. Payment choice also affects sample size: Data Quality vs. Quantity for Machine Learning

In different contexts (ML technique, data domain, and possibly at different times), different strategies can be more cost-effective for model learning: **Either compiling a larger sample of lower cost /quality labeling** or **smaller sample of costlier, and higher quality labels**

Simple, non-data-driven solutions will not work (we will see evidence of this later):

- Aim for the highest labeling quality irrespective of budget: Not always costs-effective; Nor is selecting the cheapest payments (e.g., to get more labeled instances).

- We need adaptive methods that can account for
  the market conditions (tradeoff between payment and quality),
  learning algorithm,
  data domain, and
  the learning dynamics.

Our Challenge:
Shopping for Human Labels in online labor markets

Given a budget, ML technique and data domain (a predictive task):

A machine learner that decides what payments to offer on the market for human labeling, so as to induce a model with the best predictive performance from the acquired data and a given budget.

Our Setting

How much to pay per label for the next batch of labels?

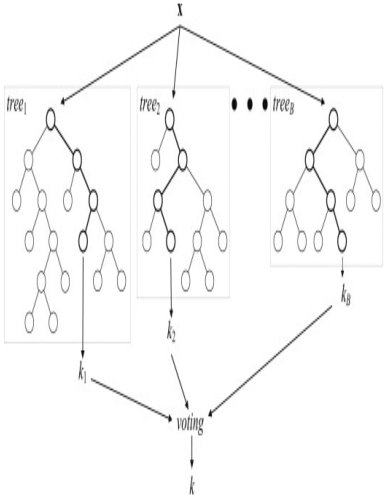
Select payment p

Request labels

Unlabeled instances **Labels**

47	100	27	81	57	37	1
0	89	27	100	42	75	1
0	57	31	68	72	90	0
0	100	7	92	5	68	0
0	67	49	83	100	100	1
100	100	88	99	49	74	1
0	100	3	72	26	35	1
0	39	2	62	11	5	
13	89	12	50	72	38	1
57	100	22	72	0	31	1
74	87	31	100	0	69	0
48	96	62	65	88	27	0
100	100	72	99	36	78	1
91	74	54	100	0	87	1
0	85	38	100	81	88	1
35	76	57	100	100	92	0
50	84	66	100	75	75	0
99	80	63	100	25	76	1
24	66	43	100	59	65	1
0	73	19	99	72	100	1
12	77	20	62	78	40	1
0	46	49	64	78	87	1
10	86	34	66	68	34	1
73	62	53	100	0	72	
54	100	34	75	6	43	
11	100	0	69	15	43	
36	92	7	83	0	37	
46	100	10	83	34	64	
61	59	58	100	0	84	
100	84	31	100	0	88	

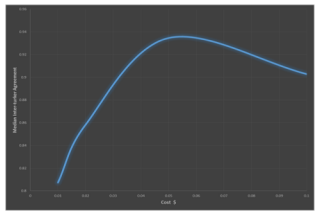
Induce a model



A set of instances are offered for labeling on the market at payment p

Market

Acquired labels added to data



Prevailing Payment/Quality Tradeoff (Unknown)

Our Approach: Adaptive Labeling Payments (ALP)

We propose **a sequential, adaptive** selection of payments for labeling:

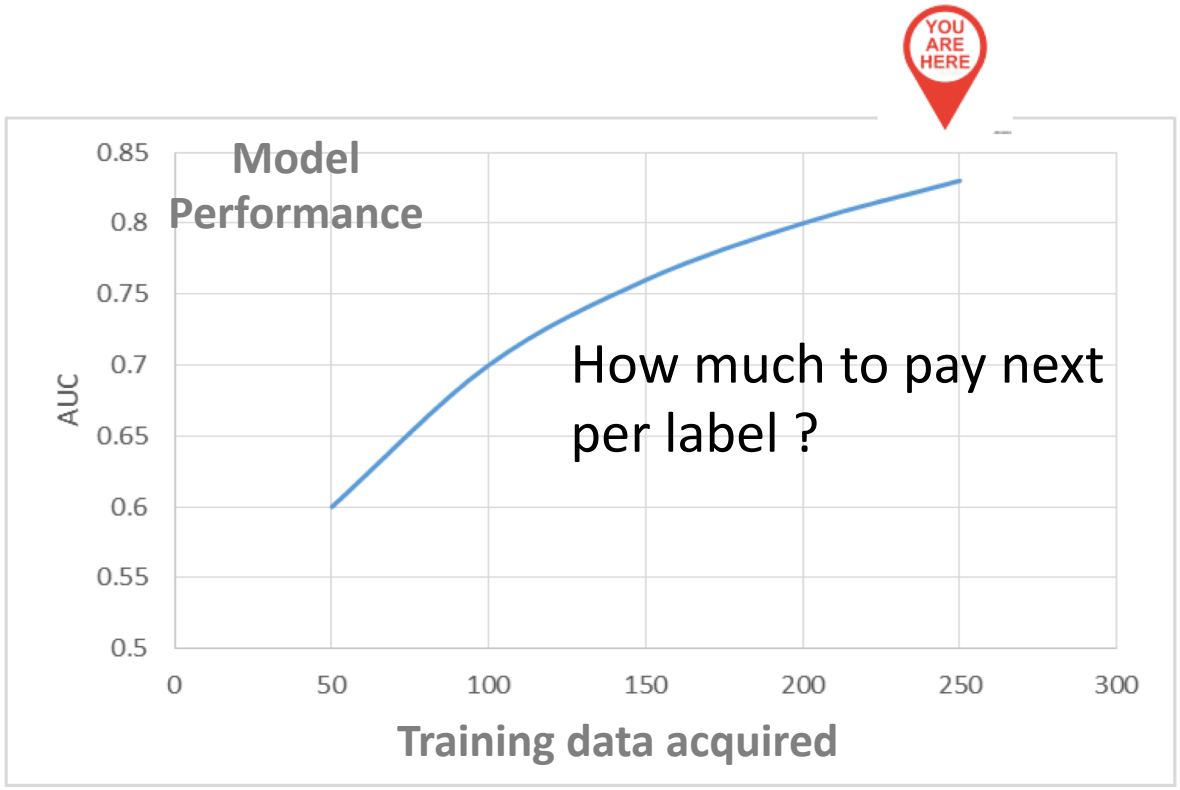
At each step, dispense a portion of our budget, $b << budget$, to buy labels at payment p^i

Select payments $p^1, p^2, \ldots p^I$ to acquire a labeled set L , such that:

$$\arg\max_{\{p^1,p^2,\ldots,p^I\}} Performance(M(L))$$

subject to $\sum_{i=1}^{I} b \cdot I \leq budget$

b: budget spent at each phase
$M(L)$: Model induced with inducer M from labeled set L



The Adaptive Labeling Payment Setting

ALP policy: What payment per label to offer next?

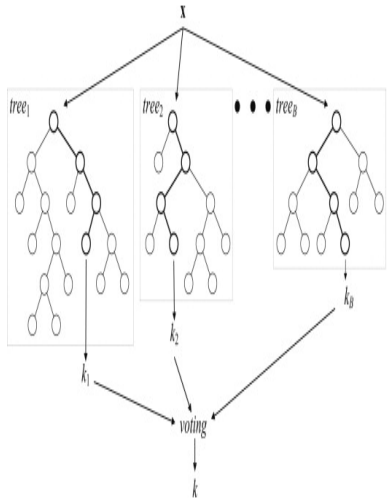
pay p

Request labels

Unlabeled instances **Labels**

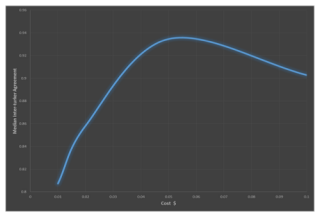
47	100	27	81	57	37	1
0	89	27	100	42	75	1
0	57	31	68	72	90	0
0	100	7	92	5	68	0
0	67	49	83	100	100	1
100	100	88	99	49	74	1
0	100	3	72	26	35	1
0	39	2	62	11	5	1
13	89	12	50	72	38	1
57	100	22	72	0	31	0
74	87	31	100	0	69	0
48	96	62	65	88	27	1
100	100	72	99	36	78	1
91	74	54	100	0	87	1
0	85	38	100	81	88	0
35	76	57	100	100	92	0
50	84	66	100	75	75	1
99	80	63	100	25	76	1
24	66	43	100	59	65	1
0	73	19	99	72	100	1
12	77	20	62	78	40	0
0	46	49	64	78	87	0
10	86	34	66	68	34	1
73	62	53	100	0	72	1
54	100	34	75	6	43	1
11	100	0	69	15	43	1
36	92	7	83	0	37	1
46	100	10	83	34	64	0
61	59	58	100	0	84	0
100	84	31	100	0	88	1

Induce a model



A set of instances are offered on the market for labeling at pay p

Market



Acquired (noisy) labels added to data

Prevailing Pay/Quality Tradeoff (Unknown)

What The ALP Method Does?

ALP policy: What payment per label to offer next?

ALP selects payment estimated to yield the best improvement in performance for a given acquisition budget

payment p

Request labels

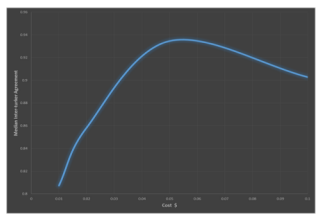
Unlabeled instances

Labels

47	100	27	81	57	37	1
0	89	27	100	42	75	1
0	57	31	68	72	90	0
0	100	7	92	5	68	0
0	67	49	83	100	100	1
100	100	88	99	49	74	1
0	100	3	72	26	35	1
0	39	2	62	11	5	1
13	89	12	50	72	38	1
57	100	22	72	0	31	1
74	87	31	100	0	69	1
48	96	62	65	88	27	1
100	100	72	99	36	78	1
91	74	54	100	0	87	1
0	85	38	100	81	88	0
35	76	57	100	100	92	0
50	84	66	100	75	75	1
99	80	63	100	25	76	1
24	66	43	100	59	65	1
0	73	19	99	72	100	1
12	77	20	62	78	40	1
0	46	49	64	78	87	1
10	86	34	66	68	34	1
73	62	53	100	0	72	1
54	100	34	75	6	43	1
11	100	0	69	15	43	1
36	92	7	83	0	37	1
46	100	10	83	34	64	1
61	59	58	100	0	84	1
100	84	31	100	0	88	1

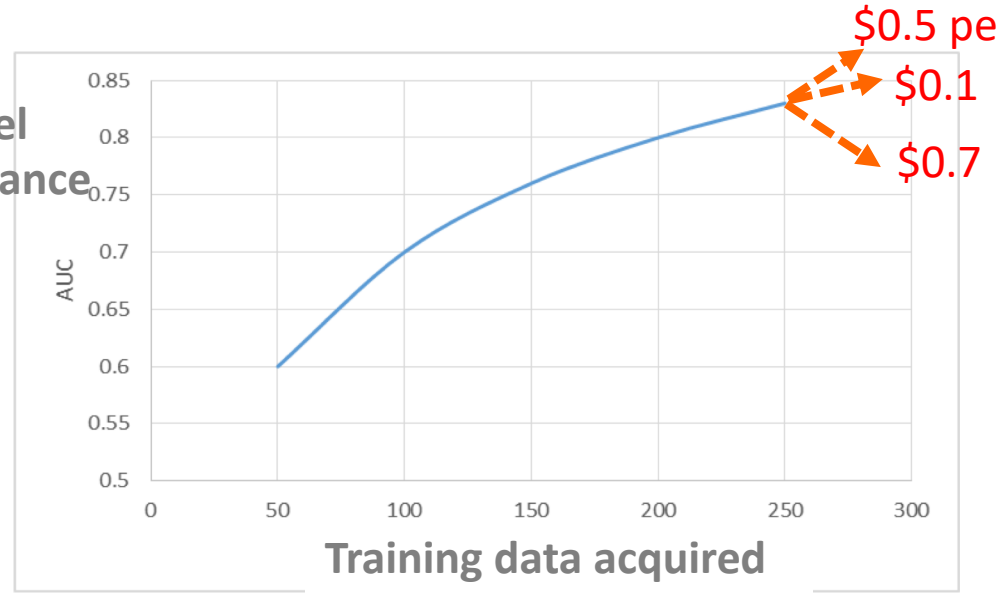
A set of instances are offered for labeling on the market at payment p

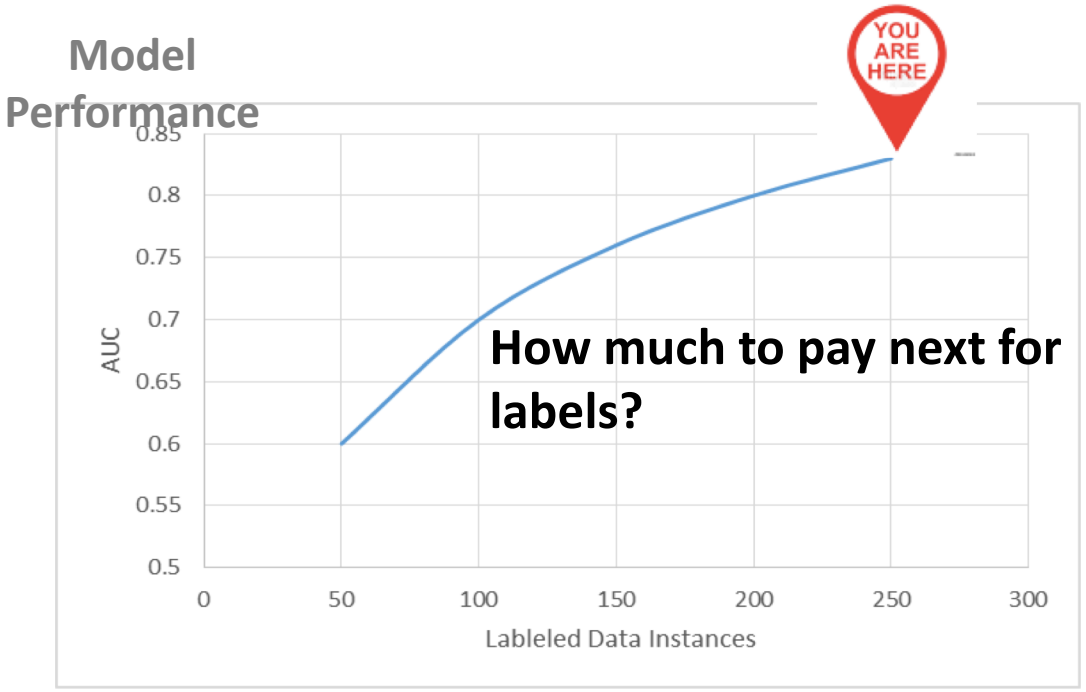
Market

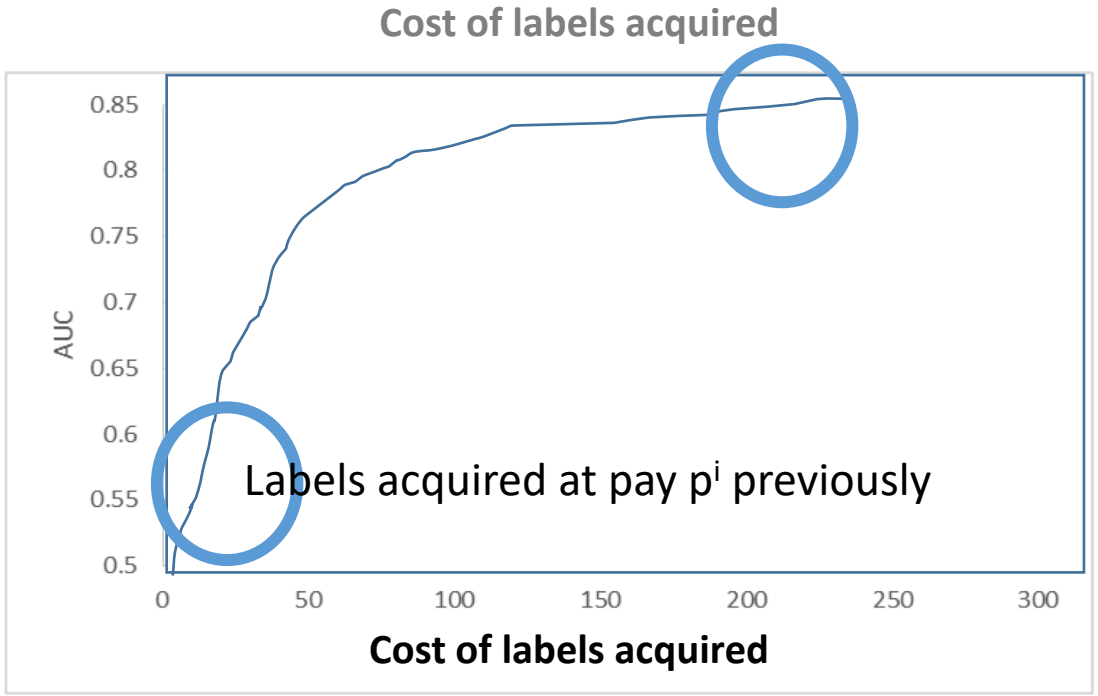


Acquired labels added to data

Prevailing Pay/Quality Tradeoff in the Mark (Unknown)







At any given time, we want to assess the effect on performance of acquiring labels at payment p^i

Q: Is a previously observed change in performance, after acquiring labels for p^i, a useful approximation of the effect on performance of the same payment now?

No!

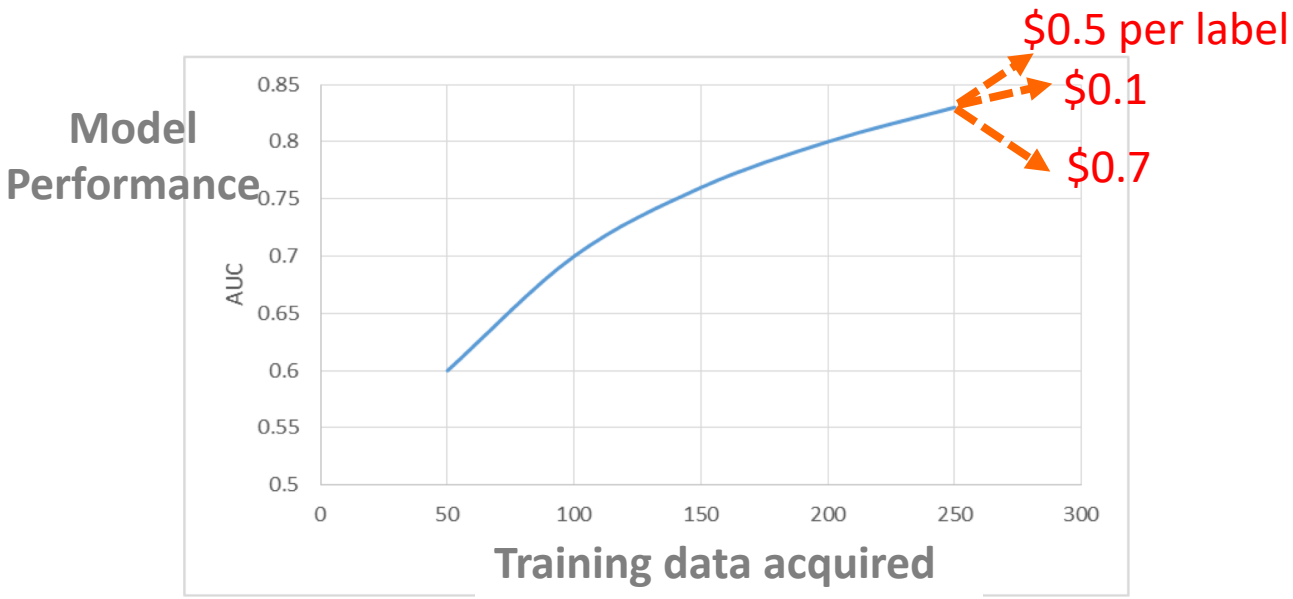
A given payment may produce different benefits to learning at different points along the learning curve

Tradeoff between payment & quality may change over time.

We saw our goal earlier: Project the improvement from different payments

Projection is very difficult: We don't know what labels will be produced by humans on the market for any given payment we may offer.

Our solution: Use our history in a clever way



We assess the impact on performance of omitting training instances (previously) labeled at a given payment

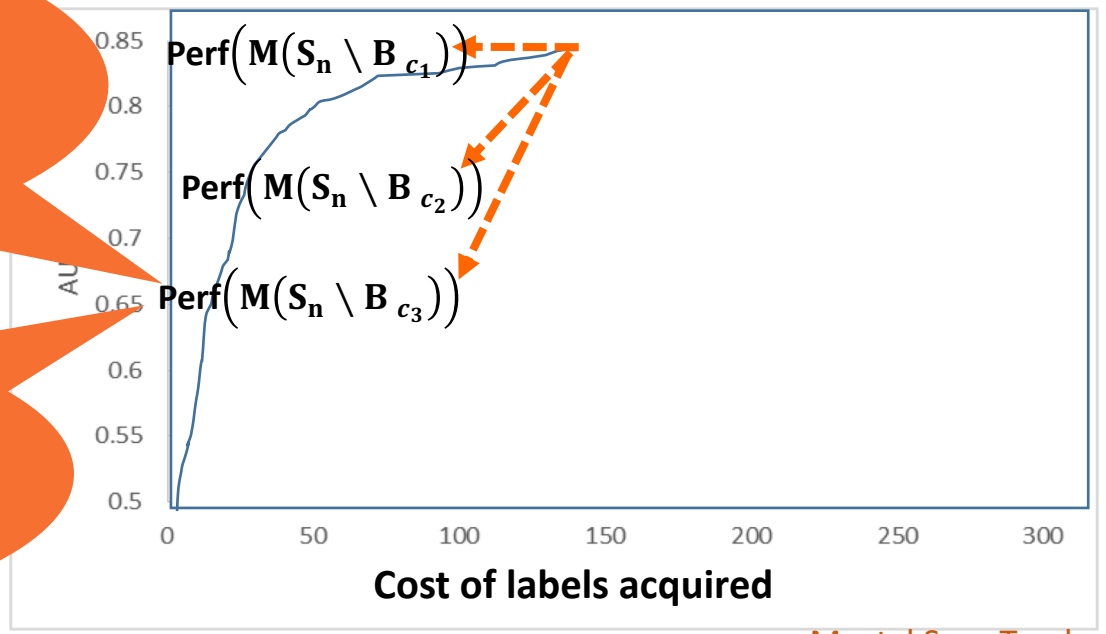
Instead of projecting: Assess the value of labels already acquired from humans for a given payment c_k in the past.

Simple idea: The loss/gain in performance after omitting instances acquired recently for c_k per label, at a total cost of *b* → A proxy of the expected improvement from paying those payments in the market now!

$$price = \arg\max_{i, b(\mathcal{D}_i)=b_0} \mathcal{R}(\mathcal{D}_{train}) - \mathcal{R}(\mathcal{D}_{train} \setminus \mathcal{D}_i)$$

Estimated performance of a model induced after omitting instances labeled at c_1 per label, and at a total cost of *b*

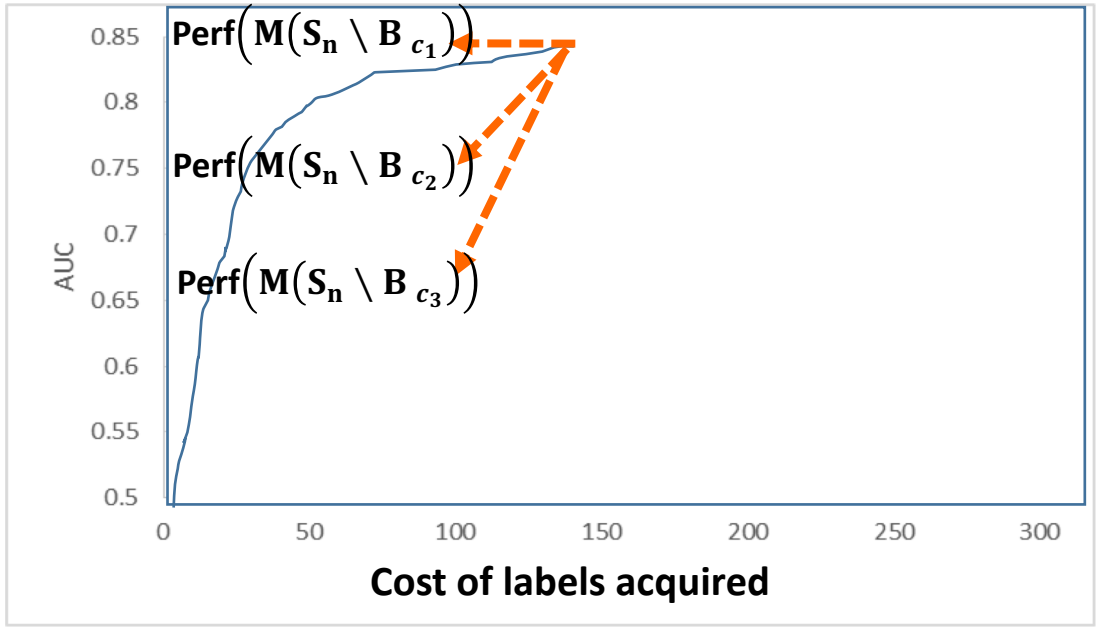
Proxy of the **benefit to** from buying human labels now at payment c_k



Adaptive Labeling Payments

Key benefits of our approach:

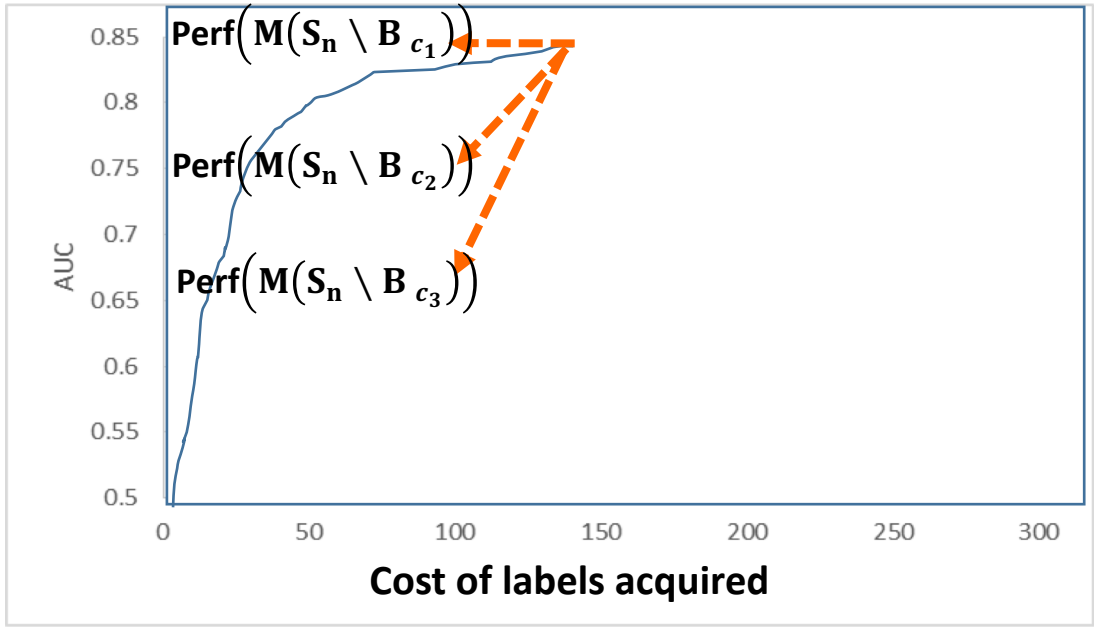
- Allow to assesses **cost-effectiveness to learning of different labeling payments** (as opposed to merely acquire the most accurate labels!)
    - Considers the **value of a payment to ML: both from the labeling <u>quality</u> and the <u>sample size</u>** that can be acquired for a given budget
- Accounts for learning dynamics: Can estimate the value of a payment **at the a given point along the learning curve, for any modeling technique and learning task.**
- **Model-agnostic:** Empirical estimation that is directly applicable with any model type.
- **State-of-the-art performance**



Can this be done?

Our approach relies on empirical estimation of model performance (before/after) omitting instances.

Is there hope to using noisy labels to assess which payment yields the largest drop in performance (and thus expected to yield the greatest benefits)?



Yes! This is because:

1. Classification error on noisy data is unbiased in order: we can correctly infer which payment yield the highest/lowest loss/error (Geva & Saar-Tsechansky, 2020)

2. CV on noisy data is also nearly unbiased estimator of the loss (Molinaro et al. 2005).

The correct order of models' lose/generalization accuracies can be recovered using noisy data

Let $0 \leq q \leq 1$: the probability that a label is correct for any given instance

$Acc(\mathcal{M})$: model $\mathcal{M}$'s true generalization accuracy

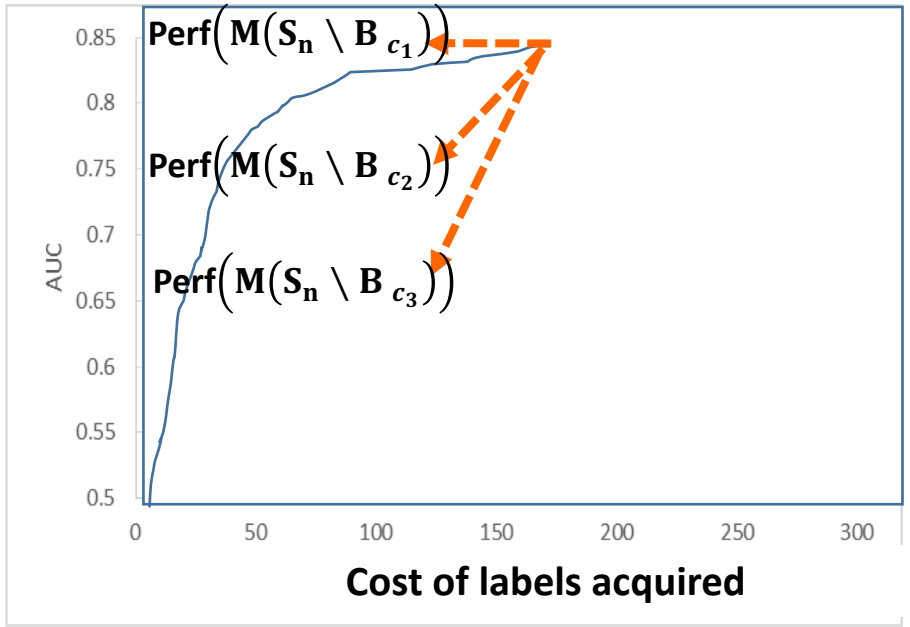
When the likelihood of prediction error for a given instance X_i is independent of whether X_i was correctly labeled:

Model $\mathcal{M}'s$ estimated accuracy measured over noisy test data with labeling accuracy q is:

$$(1).\ M\_Acc(\mathcal{M}, q) = Acc(M)(2q - 1) + 1 - q$$

From (1) : if $q > 0.5$ (labels are more often correct), then

$$(2)\ \forall\ Acc(\mathcal{M}_a) > Acc(\mathcal{M}_b) \Leftrightarrow M\_Acc(\mathcal{M}_a, q) > M\_Acc(\mathcal{M}_b, q)$$



Related Work

To our knowledge, ours is the first work to propose a framework to **prescribe labeling payments** in real-world, online labor markets

- Buying labels and selecting "oracles" with known quality for different labeling costs (Yang and Carbonell, 2012)
  - Not applicable to a labor market setting: assumes one can request labels at a desirable quality from an oracle.
- Repeated labeling using multiple noisy labelers (Ipeirotis et al., 2013)
  - Goal is to improve labeling quality irrespective of cost. No prescription of what to pay per label: assumes predetermined (given) and fixed labeling cost and quality
- Design incentives for crowd workers to achieve desired average labeling accuracy (Wang et al., 2013)
  - Similar to repeated labeling: aims to improve labeling quality irrespective of effect on model performance of different cost/quality tradeoffs

ALP vs. Other Problems Addressed in Prior Work:

Active Learning
Considers which training instances to acquire labels for ?
Thus aims to estimate the differential effect on learning of different training instances

Does not consider what to pay for labels, nor that labels can be acquired for different qualities.

Active learning can possibly be combined with ALP:

Simultaneously decide which instances to label, and how much to pay for labels.

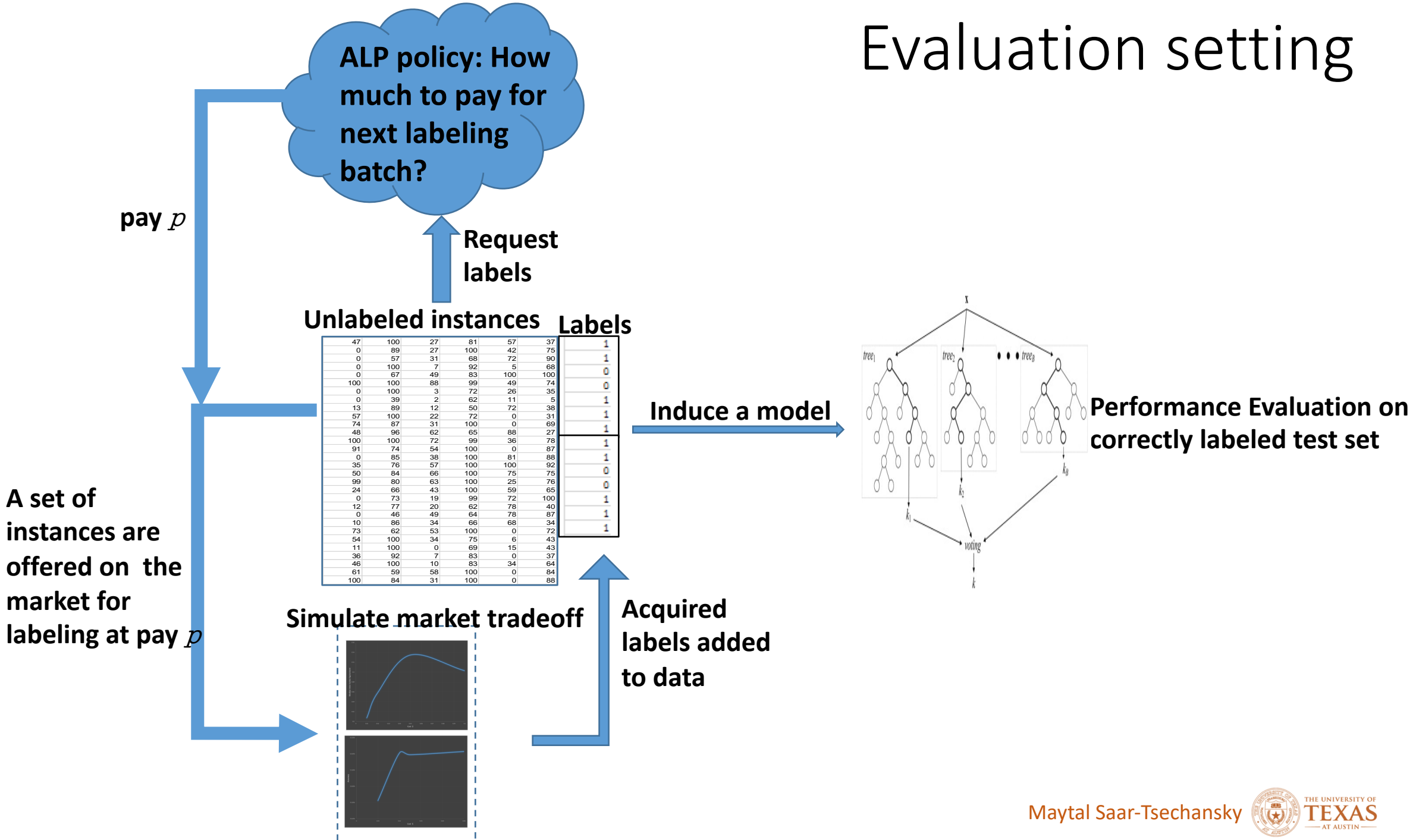
Let’s do the numbers!

Our Empirical Evaluations first aim to answer:

Does our approach acquire human labels that yield better performance for a given budget?

Is our approach’s performance robust across settings?

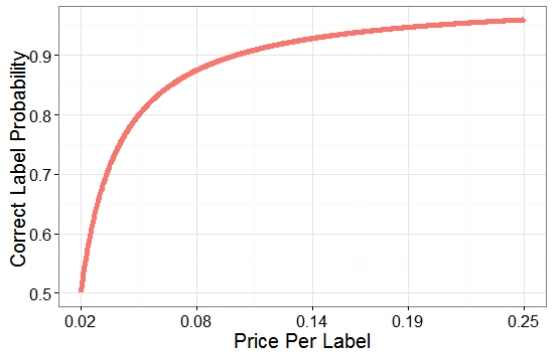
Evaluation setting



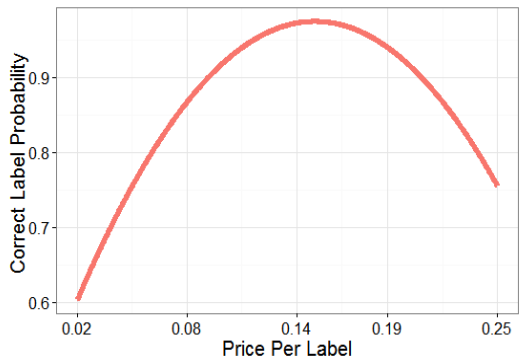
Evaluation setting

Empirical settings: Different data domains (labeling requiring human intelligence e.g., handwriting recognition), different market tradeoffs between payment and quality:

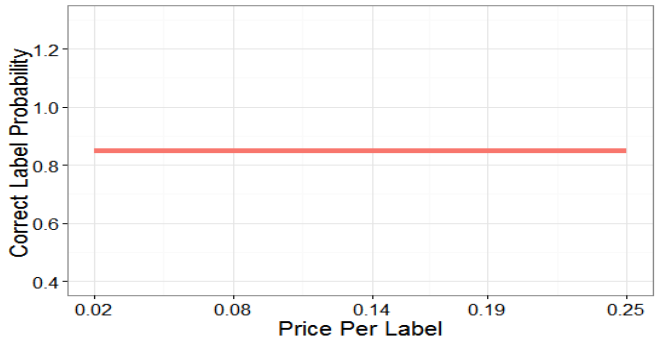
- Consider tradeoffs found in field experiments in real labor markets:



"Asymptotic"



Concave



Fixed

What are we looking for?

- **Good and consistent performance across settings**: A method can be relied on in practice if it produces consistent benefits across settings.
- **Effective for different ML model types**: Main results with Random Forest, replicated with SVM, Bagging.

Alternative approaches for selecting payments?

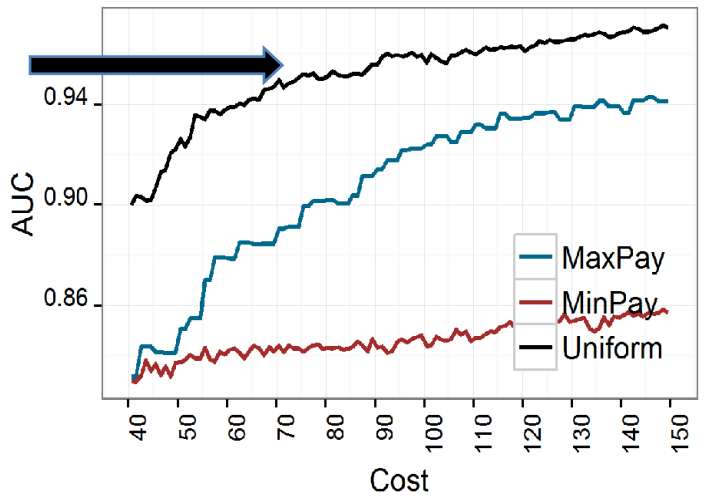
We aimed to identify a ***robust*** alternative: One that does not fail miserably under some settings:

- **MinPay**: Always select the lowest payment per label
- **MaxPay**: Always pay the highest payment per label
- **Uniform**: Uniformly draw from different payments at each batch

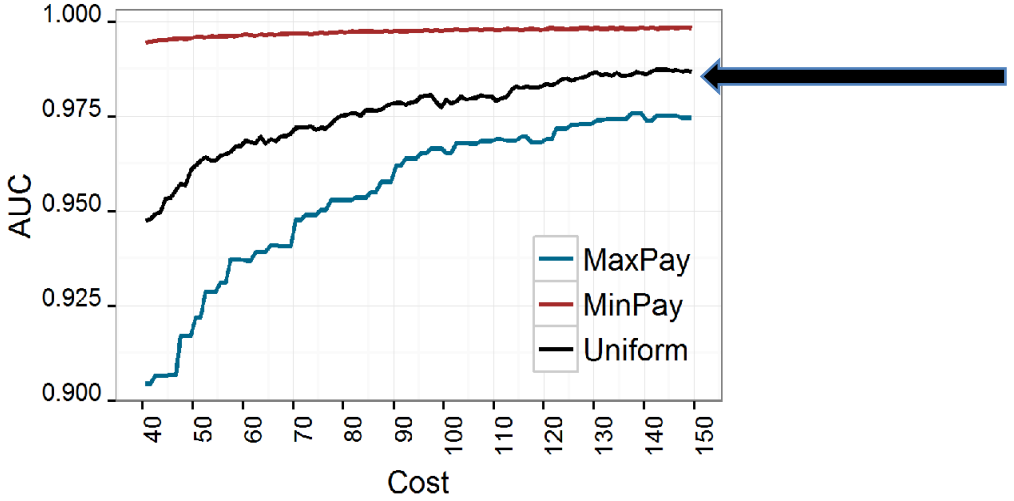
Selecting a reliable alternative

Uniform yields more robust performance across settings: It often yields good performance and is never the worst approach

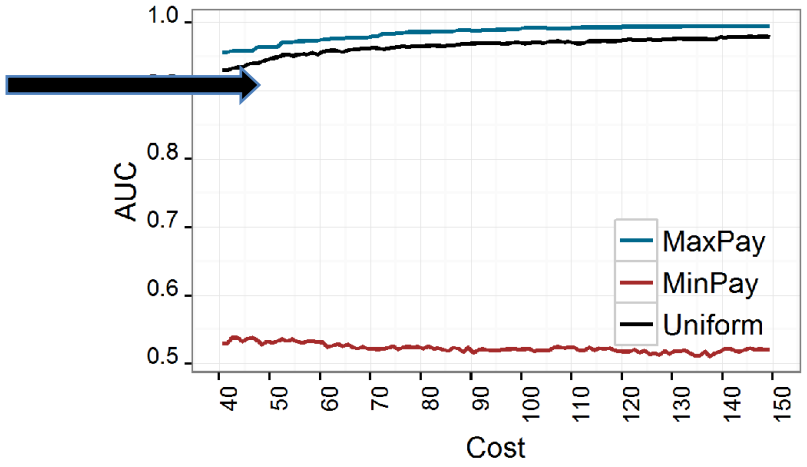
Concave tradeoff



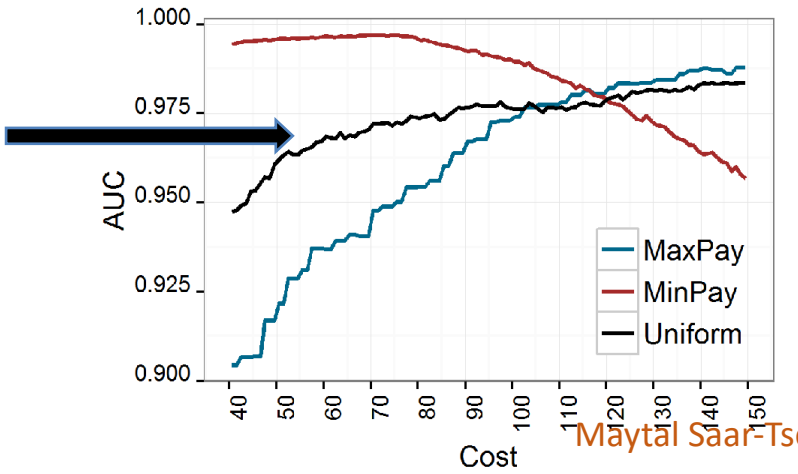
Fixed Tradeoff



Asymptotic tradeoff

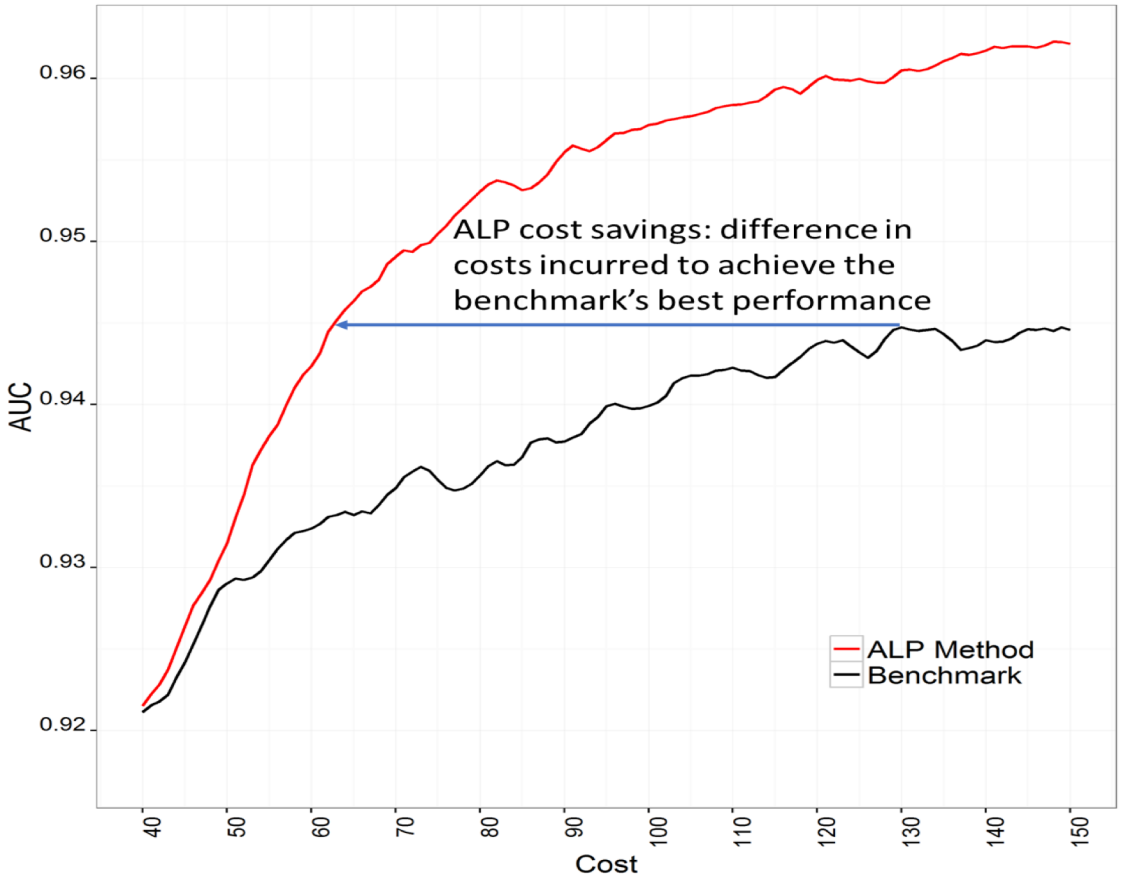


Tradeoff switches from fixed to asymptotic



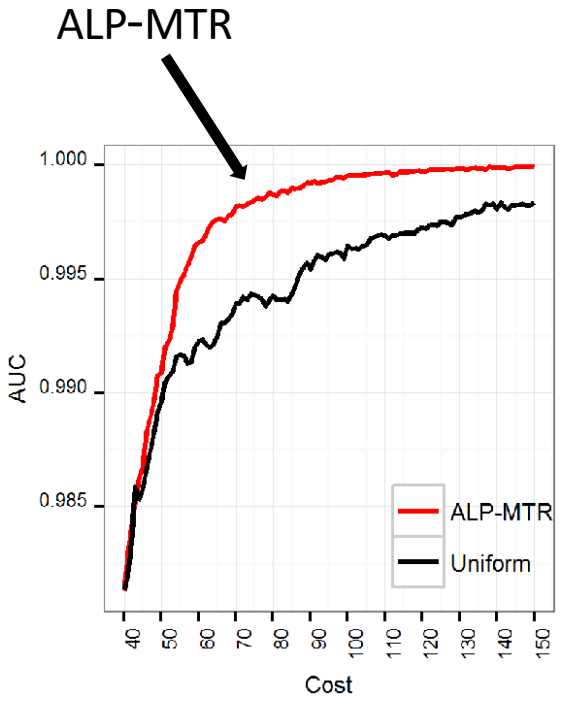
ALP's Cost Savings

Dataset	Tradeoff Function	Cost Savings enabled by ALP-MTR
Mushroom	Asymptotic	39.9%
	Concave	50.6%
	Fixed	18.4%
Spam	Asymptotic	39.7%
	Concave	60.7%
	Fixed	15.6%
Pen Digits	Asymptotic	34.3%
	Concave	38.2%
	Fixed	23.0%
Average ALP Savings		**35.6%**

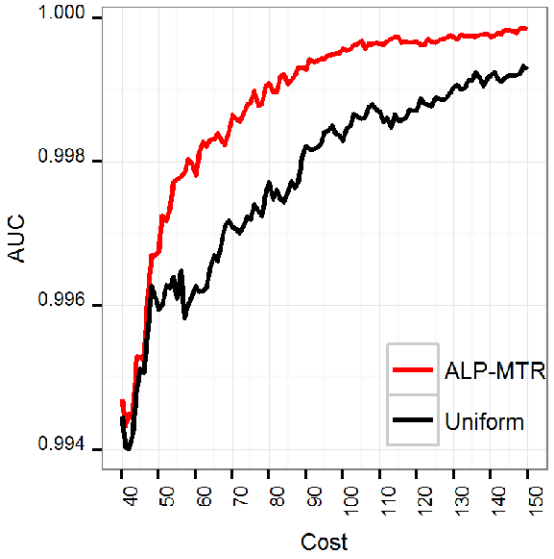


Results reflect average over 20 runs

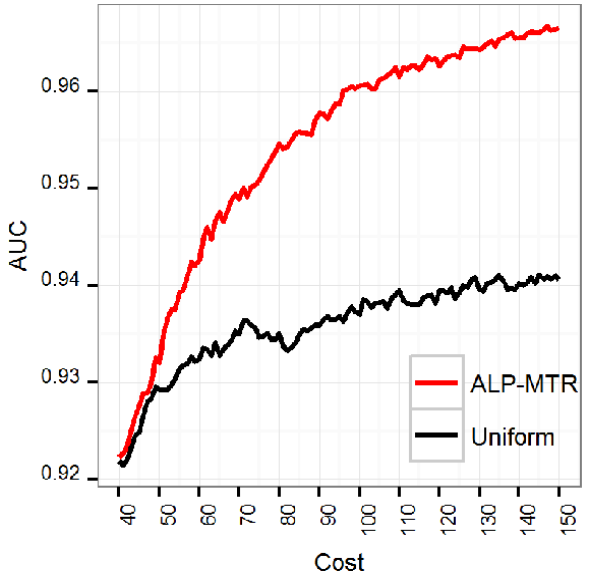
ALP vs. Uniform: When tradeoff in the market between labeling pay and quality is stationary



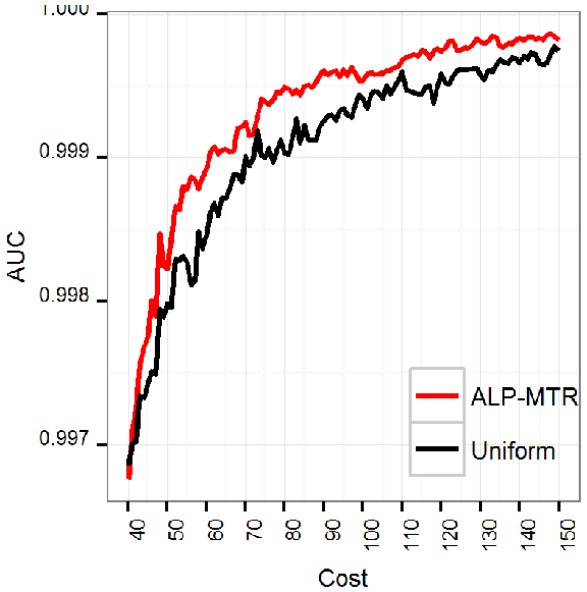
Mushroom, concave



Mushroom, asymptotic

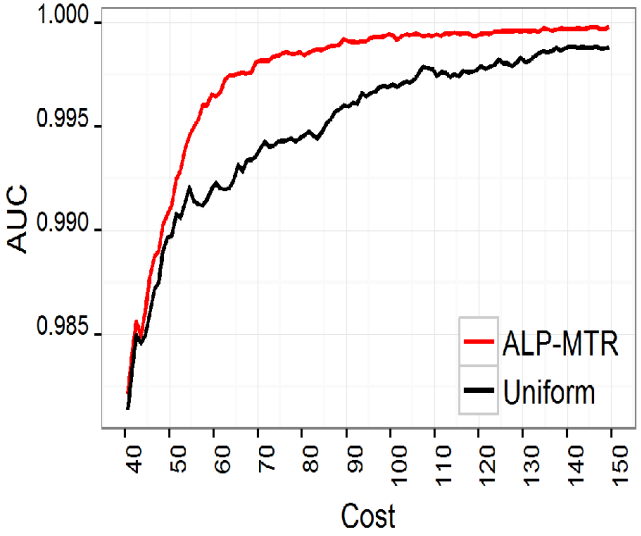


Mushroom, fixed

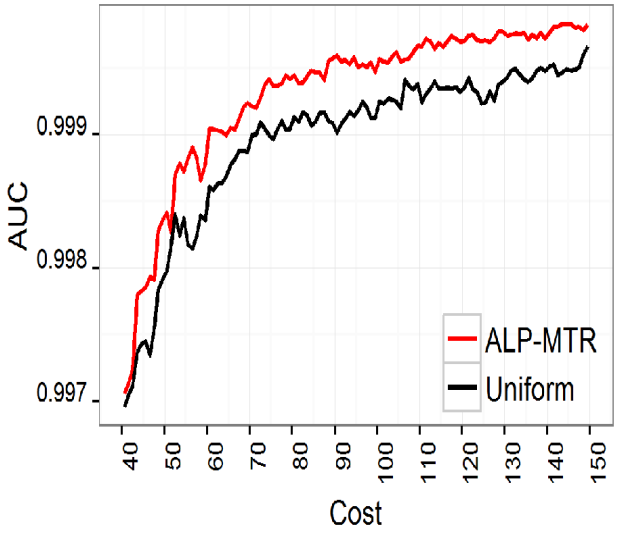


Spam, concave

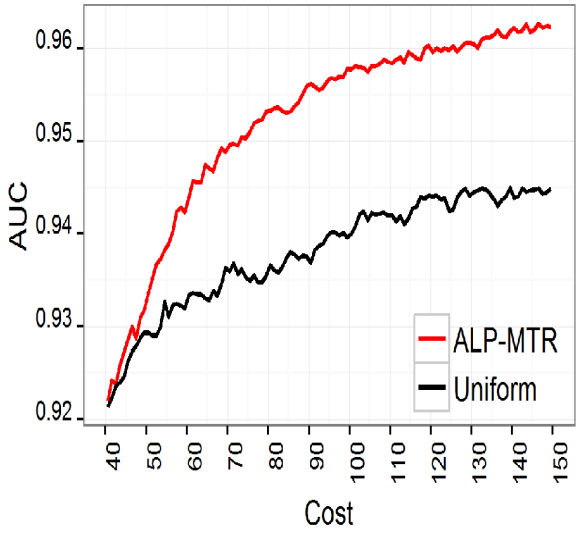
Non-stationary tradeoff between payment and quality per label in the market



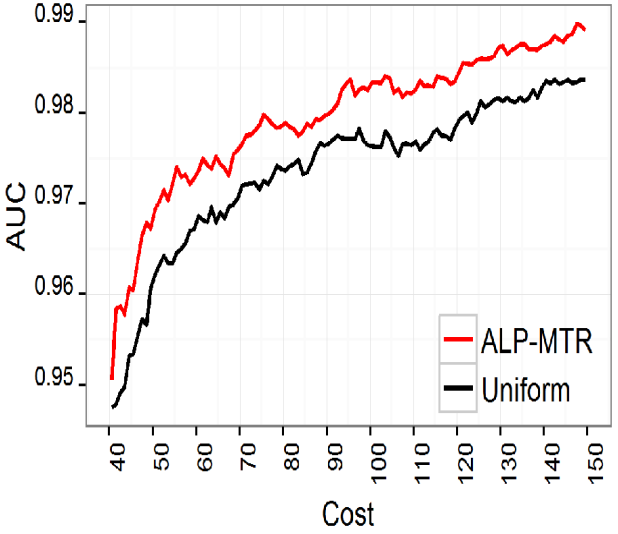
Mushroom:
Concave to Asymptotic



Mushroom:
Fixed to Asymptotic



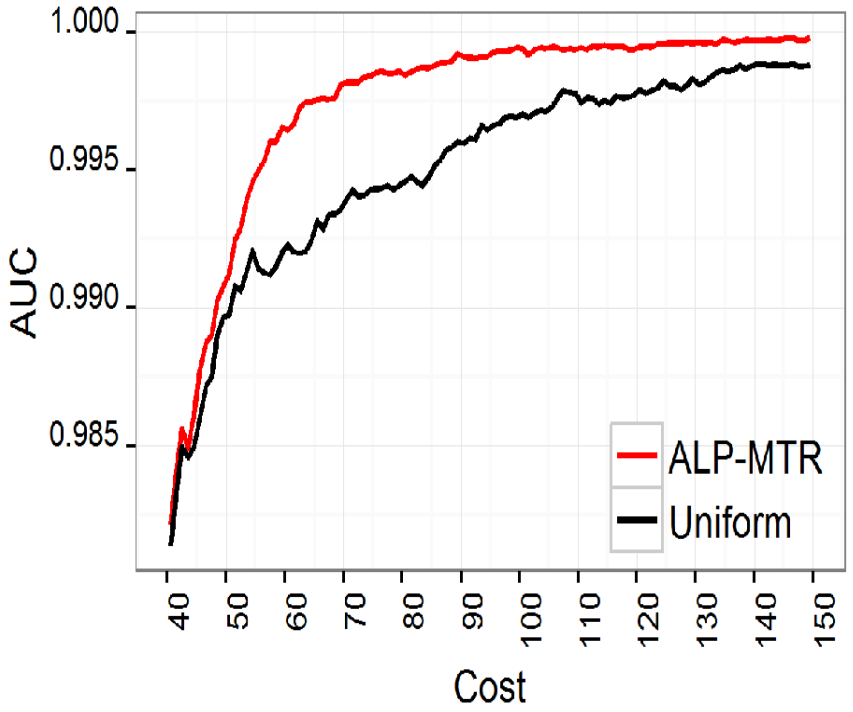
Spam:
Concave to Asymptotic



Pendigits:
Concave to asymptotic

Dataset	Change in Trade-off (From → To)	Cost Savings produced by ALP-MTR
Mushroom	Concave → Asymptotic	38.5%
	Concave → Fixed	24.3%
	Asymptotic → Concave	46.3%
	Asymptotic → Fixed	29.1%
	Fixed → Asymptotic	26.0%
	Fixed → Concave	29.9%
Spam	Concave → Asymptotic	58.9%
	Concave → Fixed	48.1%
	Asymptotic → Concave	57.3%
	Asymptotic → Fixed	25.8%
	Fixed → Asymptotic	34.6%
	Fixed → Concave	10.4%
Pen Digits	Concave → Asymptotic	29.4%
	Concave → Fixed	9.2%
	Asymptotic → Concave	38.0%
	Asymptotic → Fixed	6.2%
	Fixed → Asymptotic	31.0%
	Fixed → Concave	38.8%
Average savings:		**32.3%**

Non-stationary tradeoff between payment and quality per label in the market



A look at ALP's Labeling Quality and Quantity:

How ALP's payment "strategy" varies across settings to achieve better model performance for a given budget?

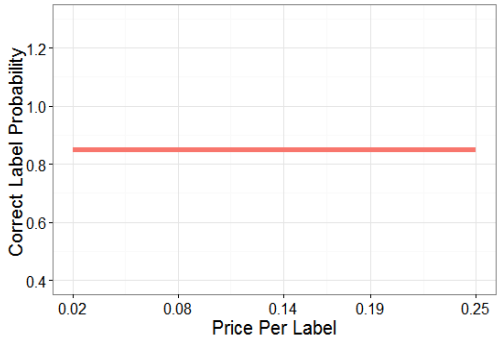
How ALP Adapts: Producing different strategies to achieve cost-effectiveness

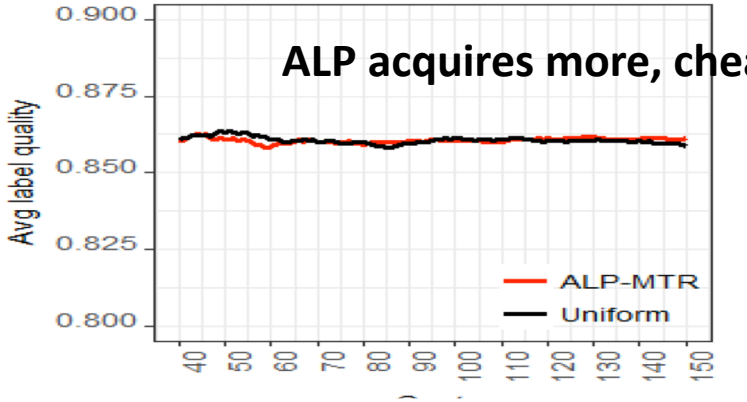
Market tradeoff between pay & quality

Labeling quality

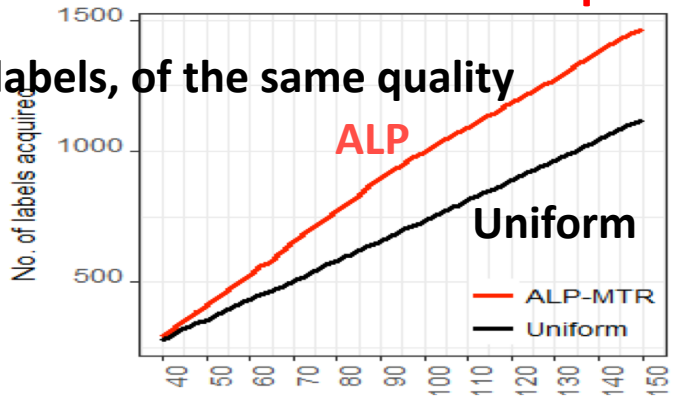
Number of labels acquired

Fixed tradeoff

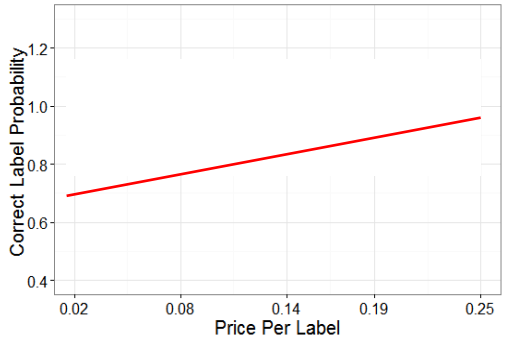


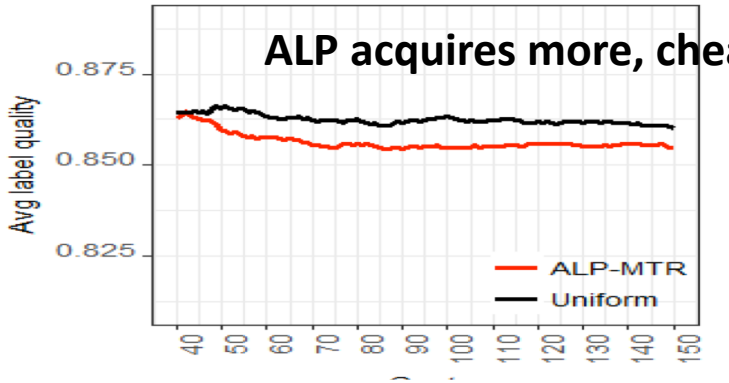


ALP acquires more, cheaper labels, of the same quality

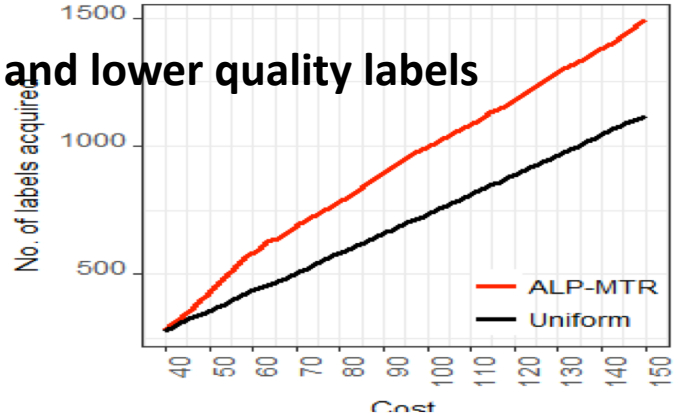


Linear tradeoff

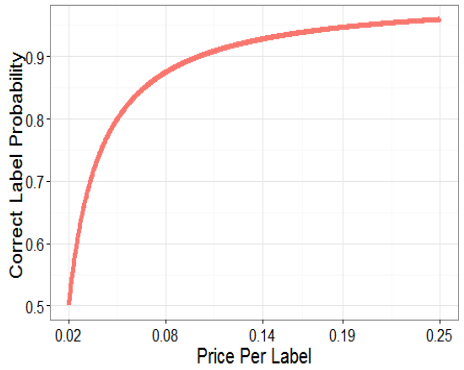


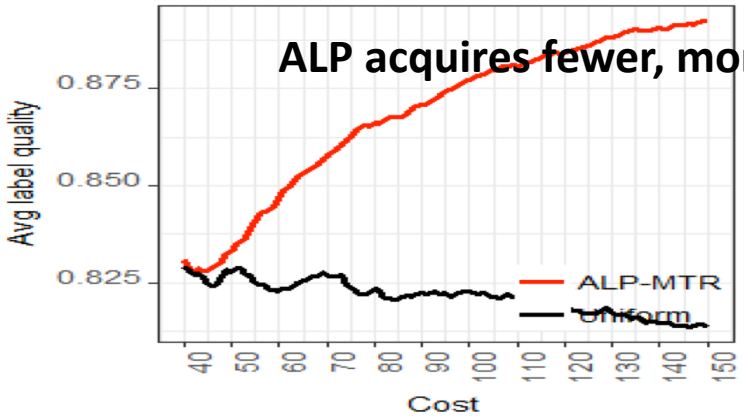


ALP acquires more, cheaper, and lower quality labels

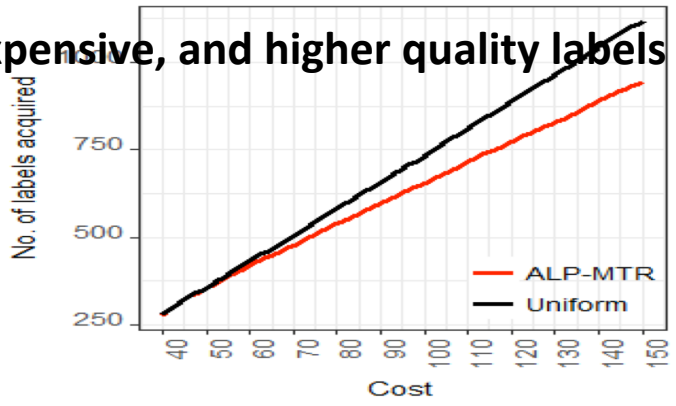


Asymptotic tradeoff

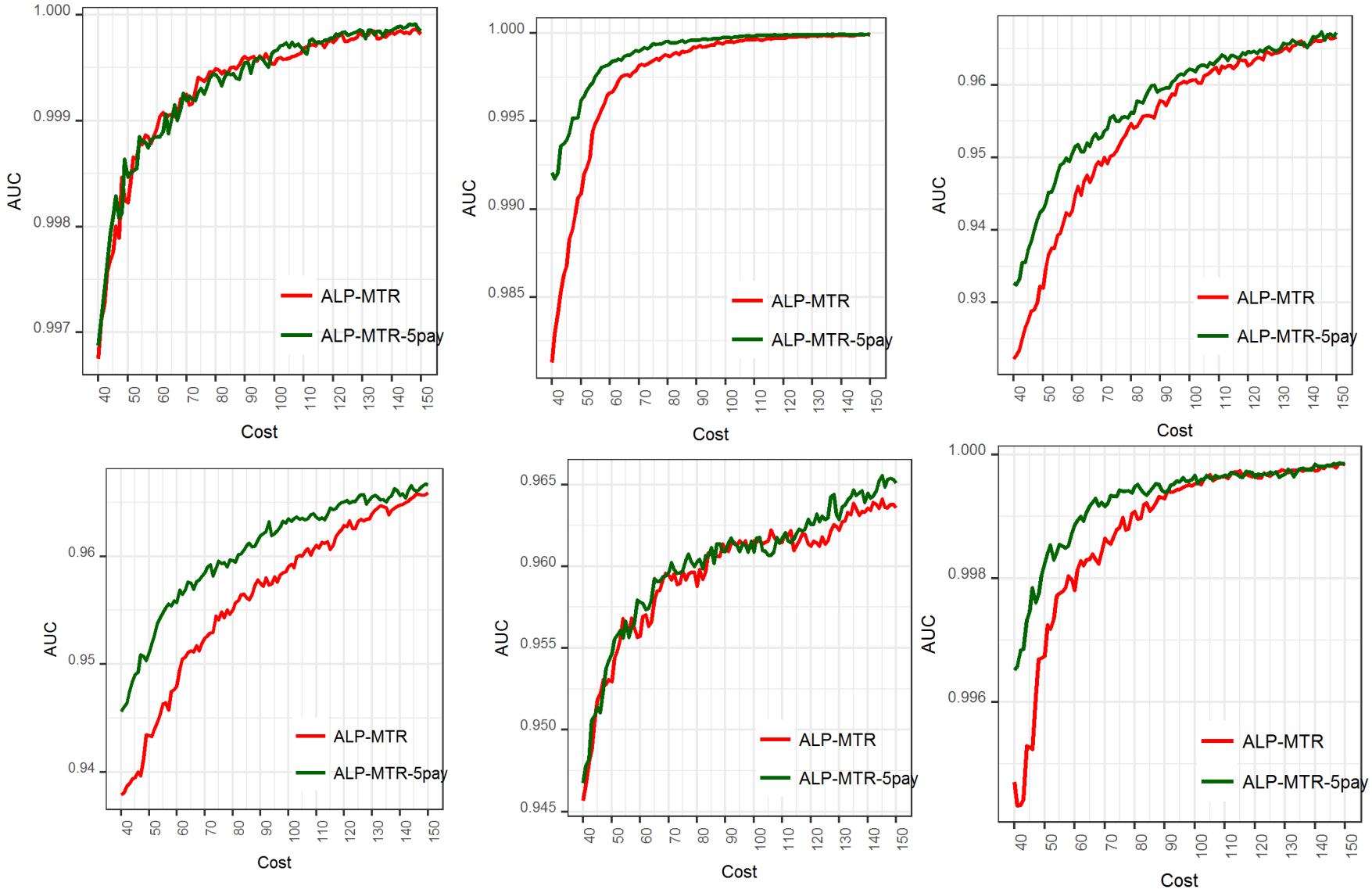




ALP acquires fewer, more expensive, and higher quality labels



Considering more payment options is beneficial



Conclusions so far

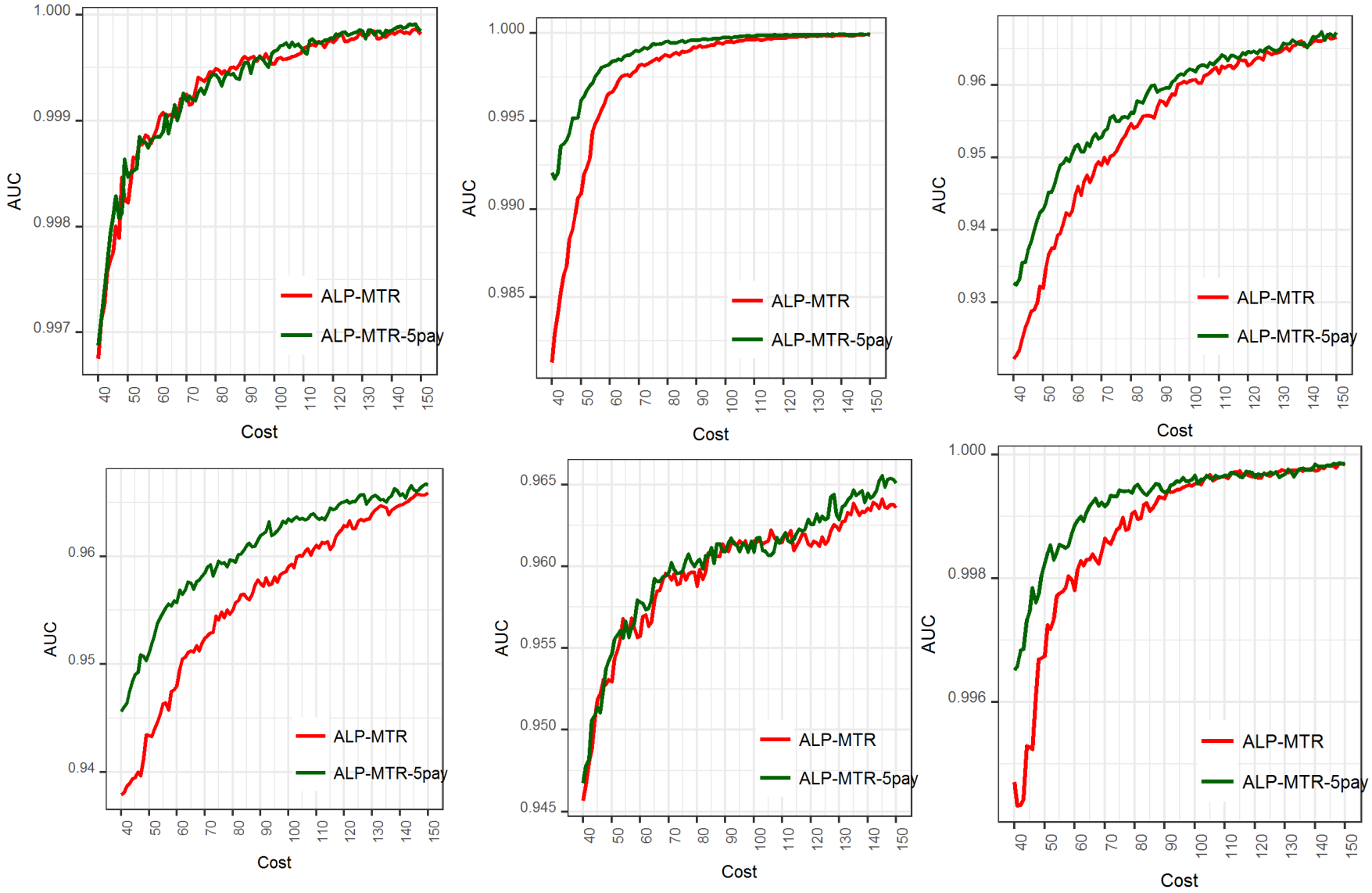
Adaptive learning of advantageous payments for imperfect human labelers :

- Our approach acquires more cost-effective human labels: achieving cost savings of more than 30%, across settings
- Consistent across market conditions: different tradeoffs between labeling cost and quality
- Consistent across problem settings: predictive task and learning algorithm.
- A generic approach: Applicable with different supervised learning algorithms ,market , and predictive tasks.

Thank you.

Maytal Saar-Tsechansky
www.maytals.com
maytal@utexas.edu

Considering more payment options is beneficial



Offering a payment if not used in t consecutive phases: Sometimes beneficial if tradeoff in the market changes