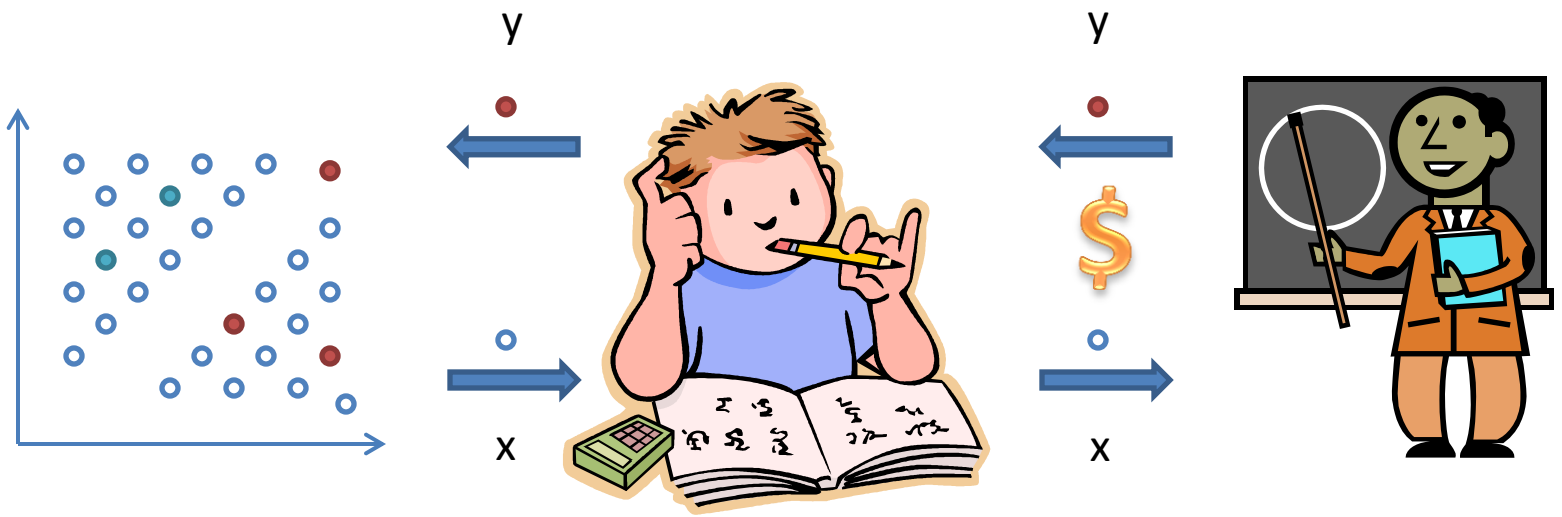
ILLINOIS INSTITUTE OF TECHNOLOGY

Towards Anytime Active Learning: Interrupting Experts to Reduce Annotation Costs

Maria E. Ramirez Loaiza
Aron Culotta
Mustafa Bilgic

Machine Learning Lab

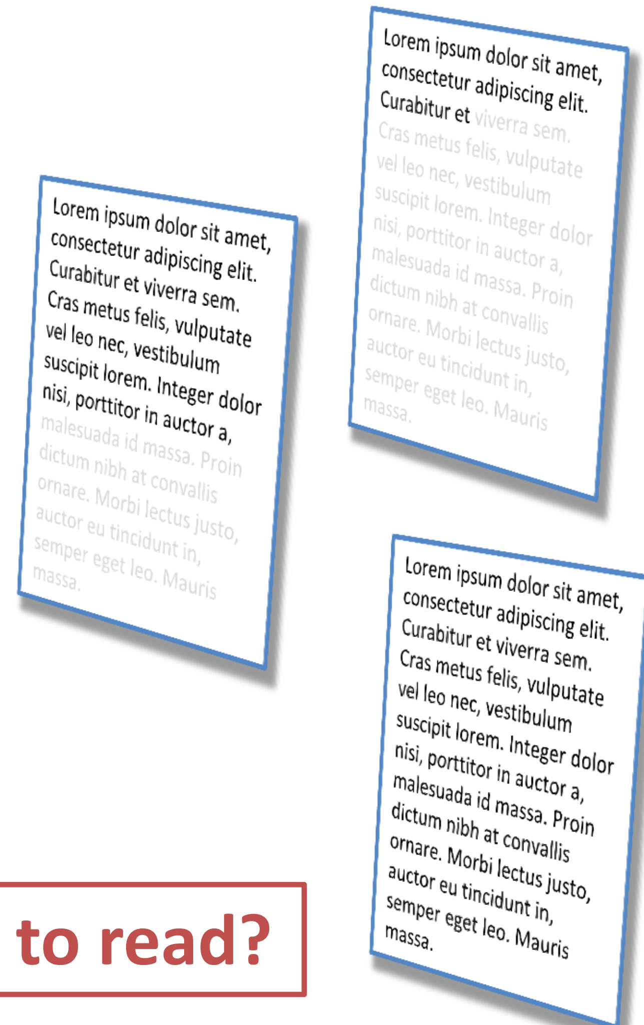
Active Learning



$$f(x) = y$$

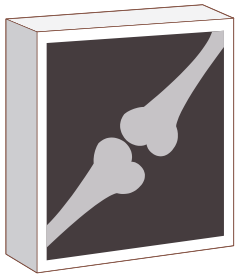
How to spend the budget?

Example: Text Classification

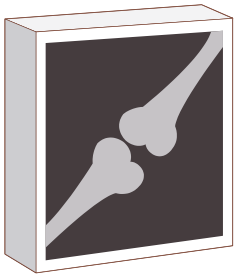


Which document and how much to read?

Example: Diagnostic Images



Example: Diagnostic Images



Example: Diagnostic Images

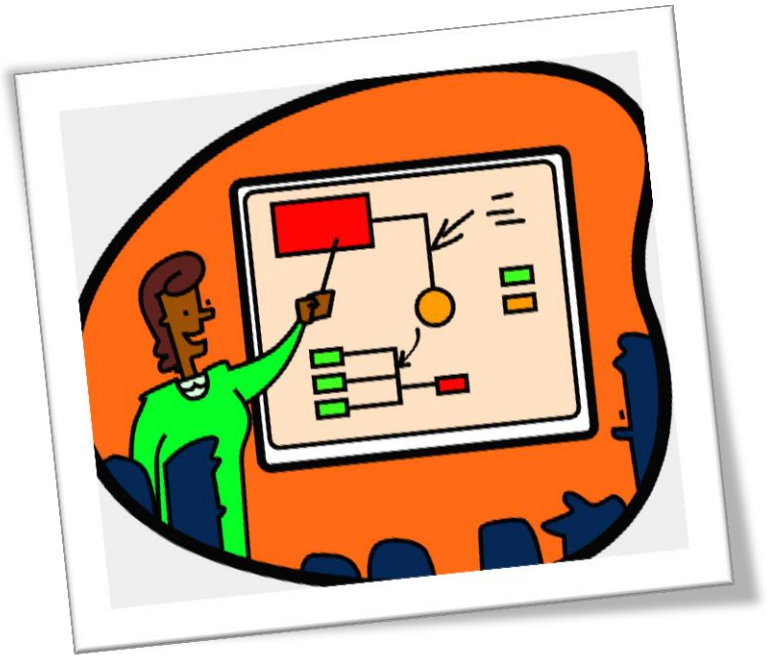
Which image and how much time to spend?

The Problem

- Anytime active learning
  - Learner interrupts the expert at anytime, ask label
- The problem:
  - Which instances to pick
  - How much time to spend on each

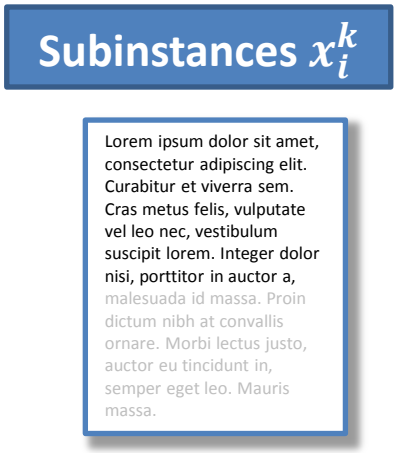
Outline

- Anytime active learning in detail
- Experimental Results
- Future Work
- Conclusion



The Problem: Formally

- x_i^k represents interruption
  - Subinstance
  - Interrupt at time k
- Value of information (VOI)



k =50 words

$$VOI(x_i^k) = Err(P_{\mathcal{L}}) - \sum_{y_j} P_{\mathcal{L}}(y_j|x_i^k) Err(P_{\mathcal{L}\cup(x_i,y_j)})$$

Current Error

Expected Future Error

Our Approach

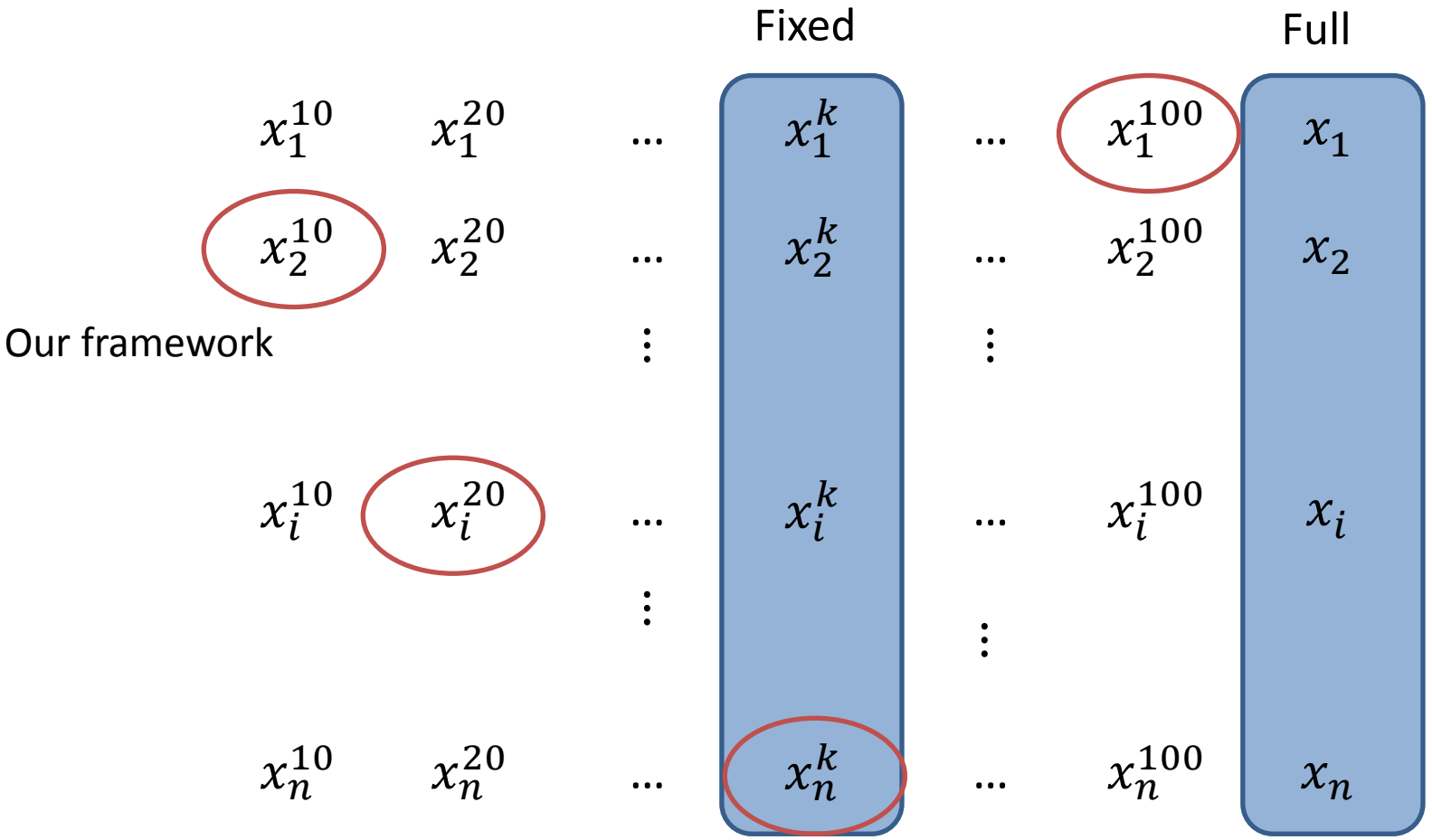
- Search over all possible subinstances
- Models tradeoff between VOI and cost

$$\arg\max_{x_i^k \subseteq x_i \in \mathcal{U}} \left(Err(P_{\mathcal{L}}) - \sum_{y_j} P_{\mathcal{L}}(y_j|x_i^k) Err(P_{\mathcal{L} \cup (x_i, y_j)}) \right) - \lambda C(x_i^k)$$

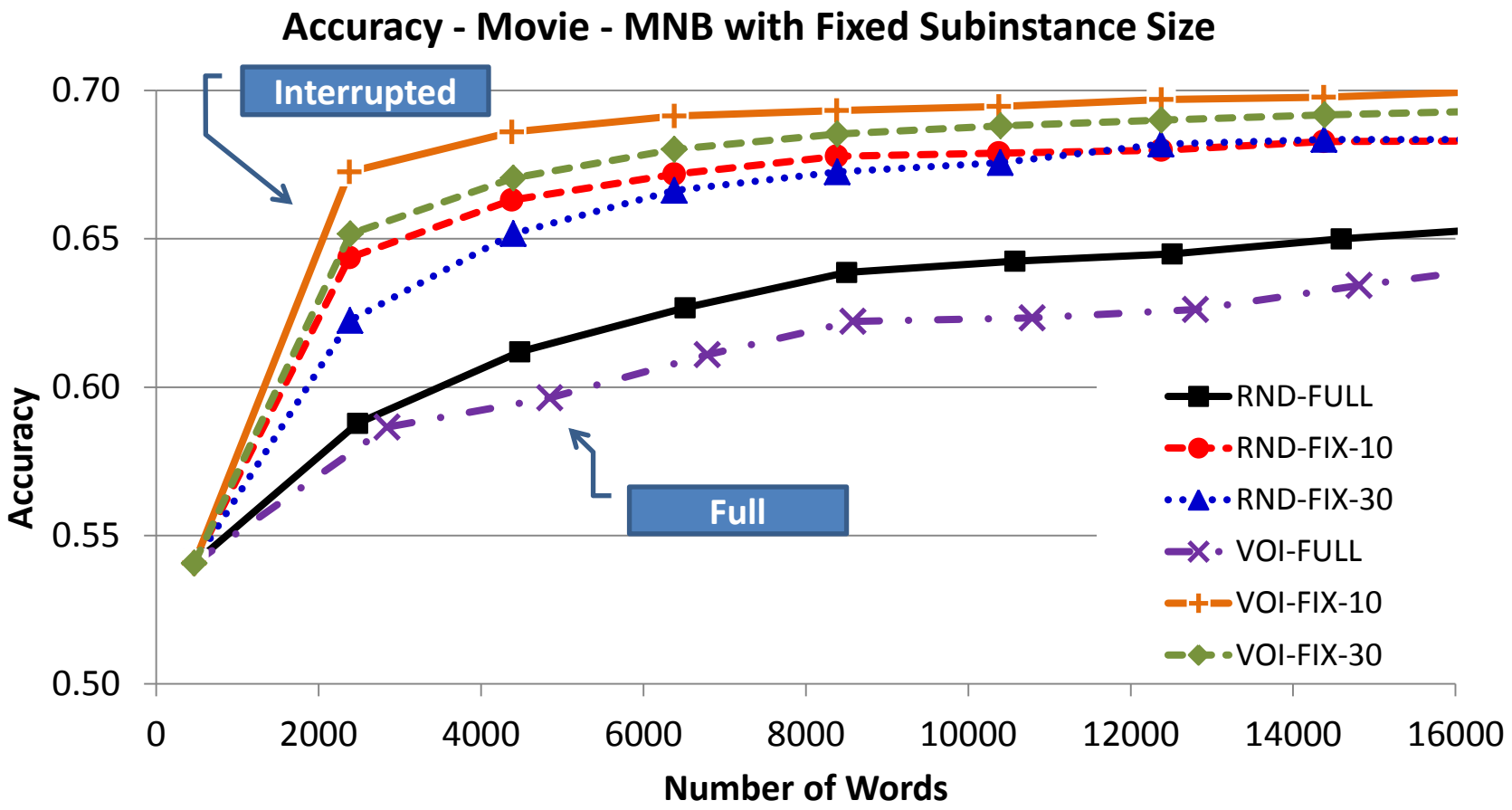
VOI of Subinstance

Labeling Cost

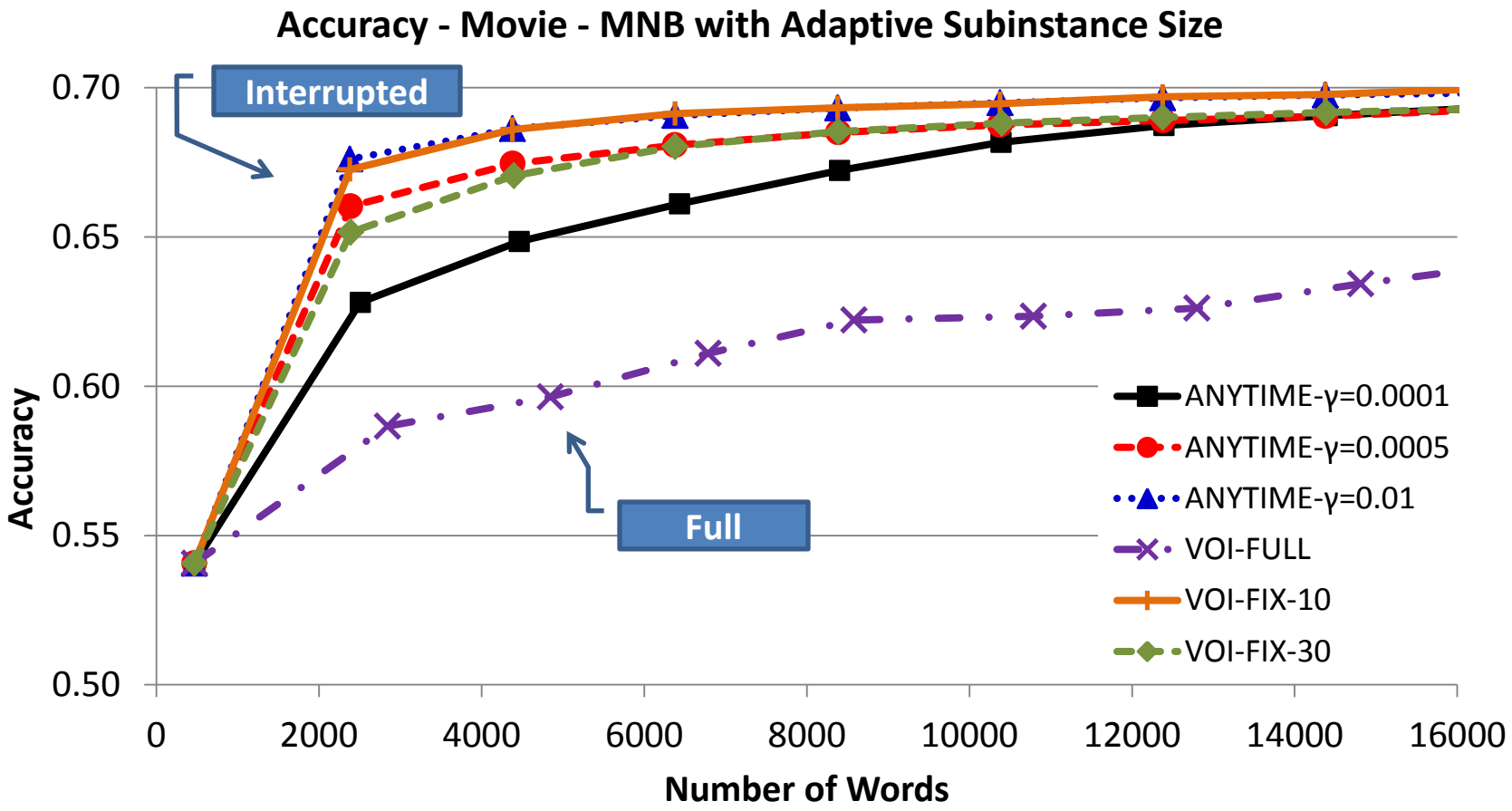
Search Space



Results – Fixed Size Interruption

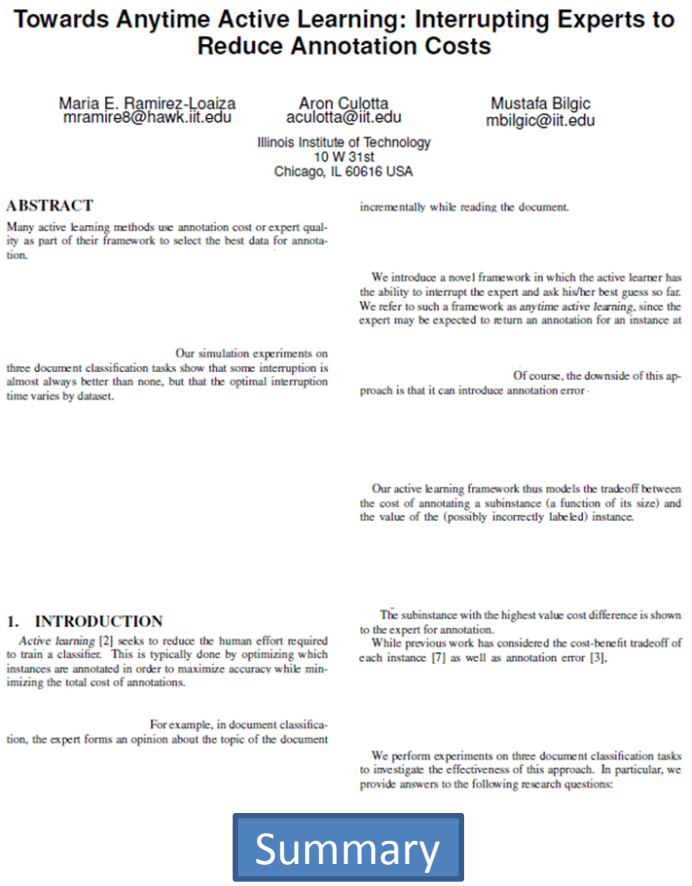


Results – Anytime Active Learning



Future Work

- Incorporate label error
- Other ways to interrupt
  - Structured reading
  - Automatic summary
- Other domains
  - e.g., vision
- User study



Towards Anytime Active Learning: Interrupting Experts to Reduce Annotation Costs

Maria E. Ramirez-Loaiza
mramire8@hawk.iit.edu

Aron Culotta
aculotta@iit.edu

Mustafa Bilgic
mbilgic@iit.edu

Illinois Institute of Technology
10 W 31st
Chicago, IL 60616 USA

ABSTRACT

Many active learning methods use annotation cost or expert quality as part of their framework to select the best data for annotation.

Our simulation experiments on three document classification tasks show that some interruption is almost always better than none, but that the optimal interruption time varies by dataset.

1. INTRODUCTION

Active learning [2] seeks to reduce the human effort required to train a classifier. This is typically done by optimizing which instances are annotated in order to maximize accuracy while minimizing the total cost of annotations.

For example, in document classification, the expert forms an opinion about the topic of the document incrementally while reading the document.

We introduce a novel framework in which the active learner has the ability to interrupt the expert and ask his/her best guess so far. We refer to such a framework as *anytime active learning*, since the expert may be expected to return an annotation for an instance at

Of course, the downside of this approach is that it can introduce annotation error.

Our active learning framework thus models the tradeoff between the cost of annotating a subinstance (a function of its size) and the value of the (possibly incorrectly labeled) instance.

The subinstance with the highest value cost difference is shown to the expert for annotation.

While previous work has considered the cost-benefit tradeoff of each instance [7] as well as annotation error [3],

We perform experiments on three document classification tasks to investigate the effectiveness of this approach. In particular, we provide answers to the following research questions:

Summary

Conclusion

- Introduce a new framework
  - Anytime active learning
- Give active learner control over expert time
- Many future directions