Machine Learning 2

DS 4420 - Spring 2020

Humans-in-the-loop

Byron C. Wallace

Today

- *Reducing annotation costs*: **active learning** and **crowdsourcing**

Efficient annotation

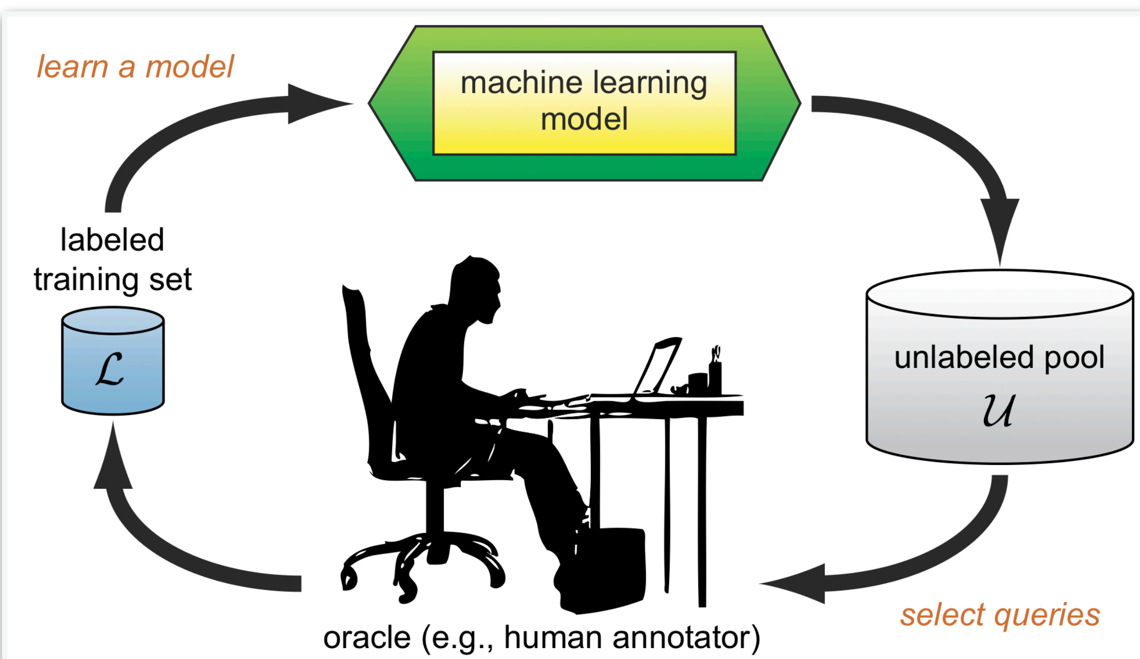
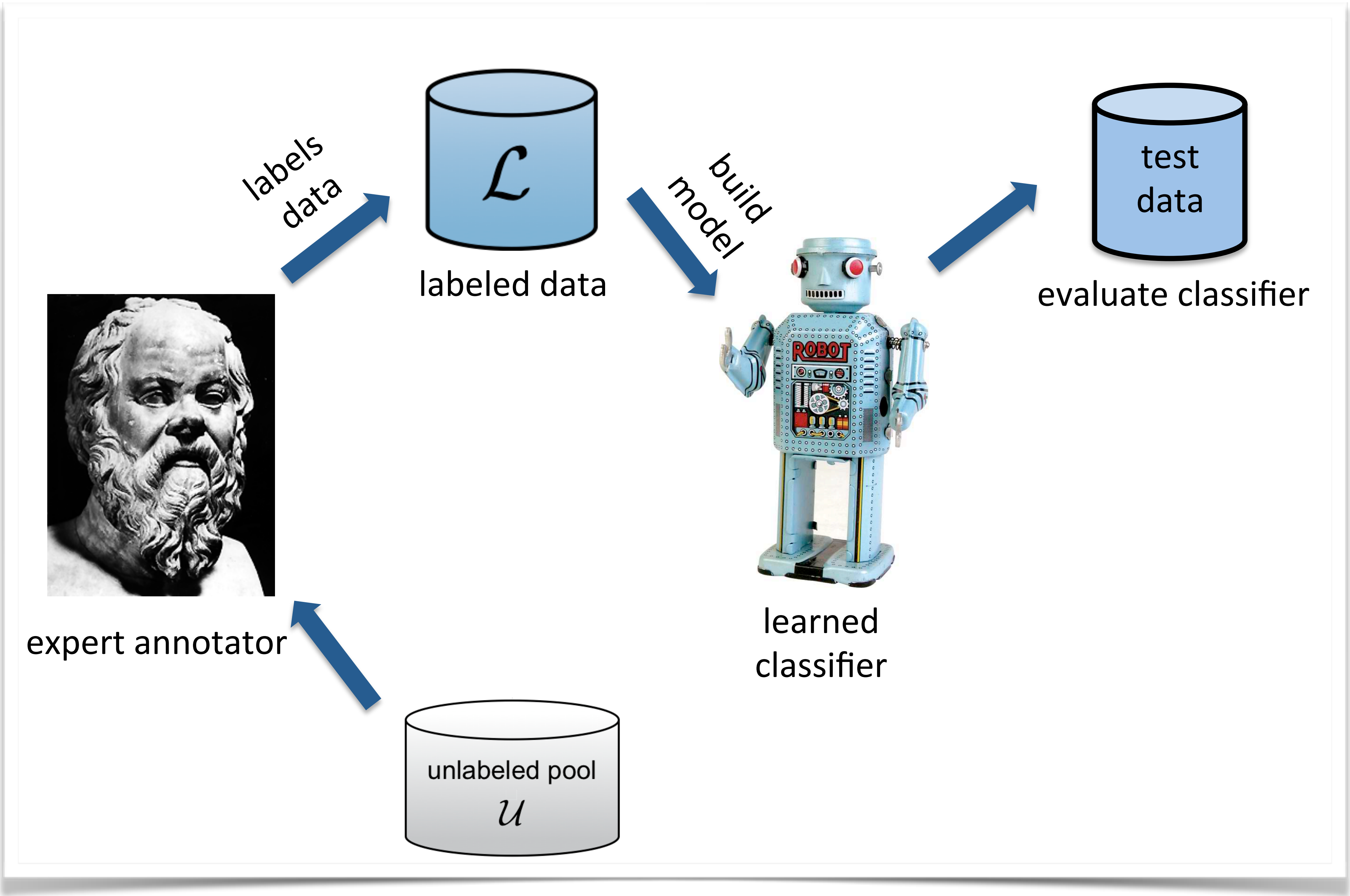


Figure from Settles, '08

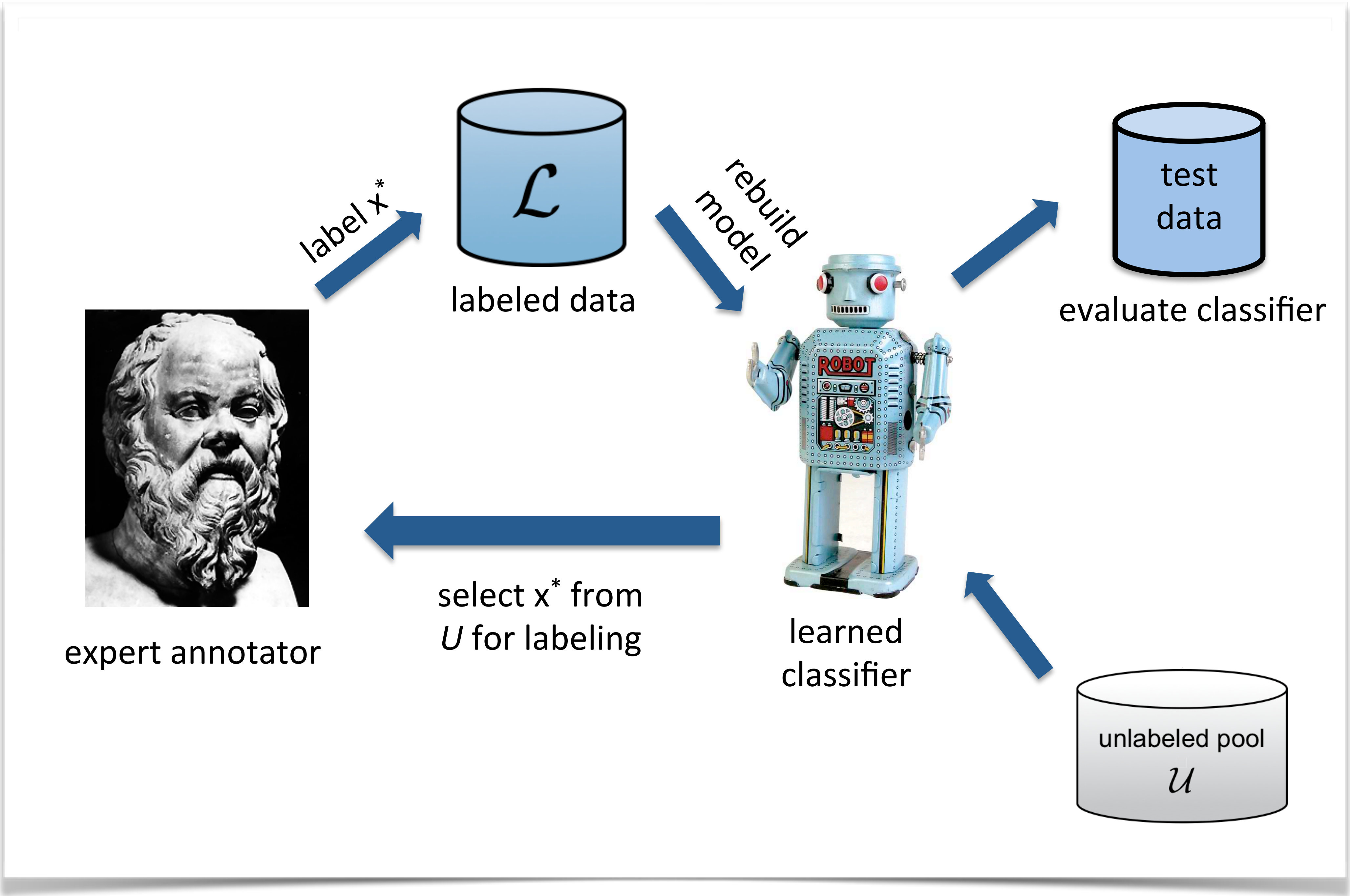
Active learning

Crowdsourcing

Standard supervised learning



Active learning



Active learning

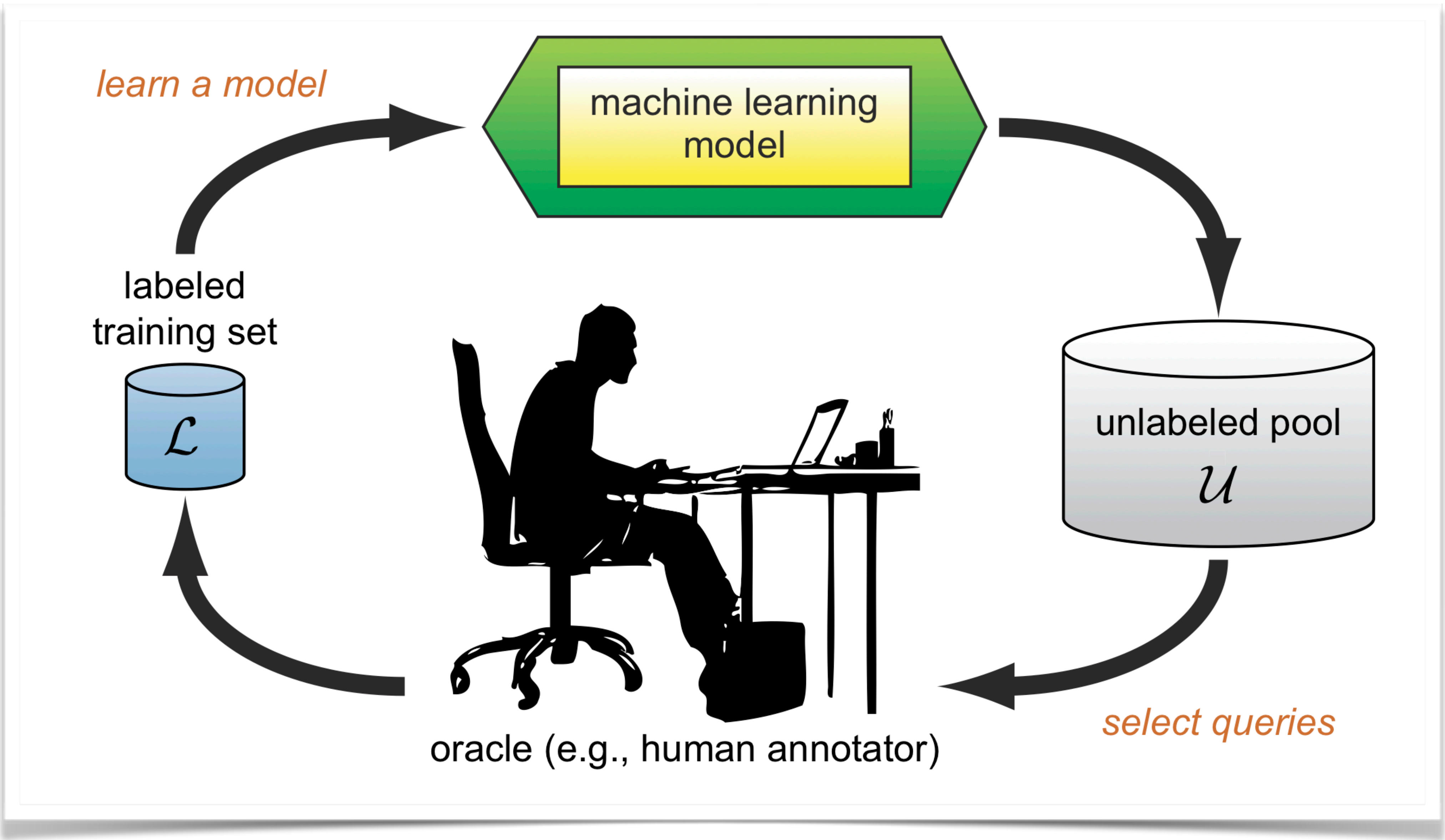
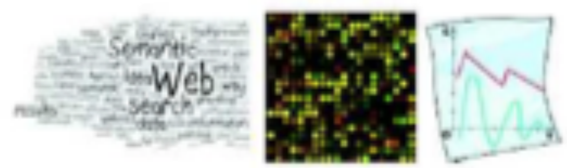
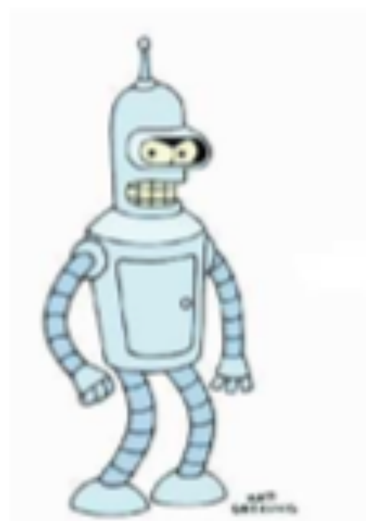


Figure from Settles, ‘08

Learning paradigms



raw unlabeled data
$x_1, x_2, x_3, \ldots$

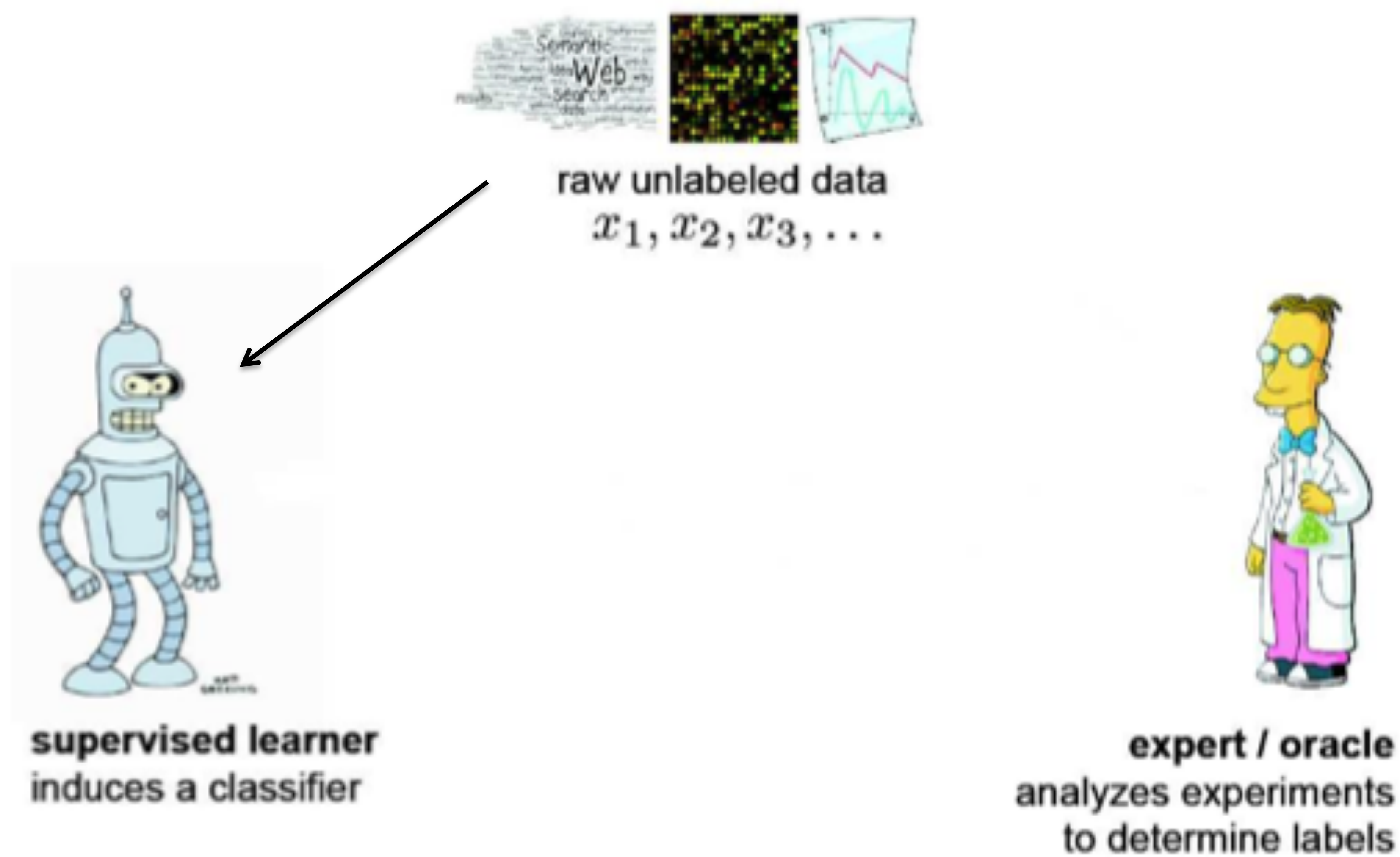


supervised learner
induces a classifier

expert / oracle
analyzes experiments
to determine labels

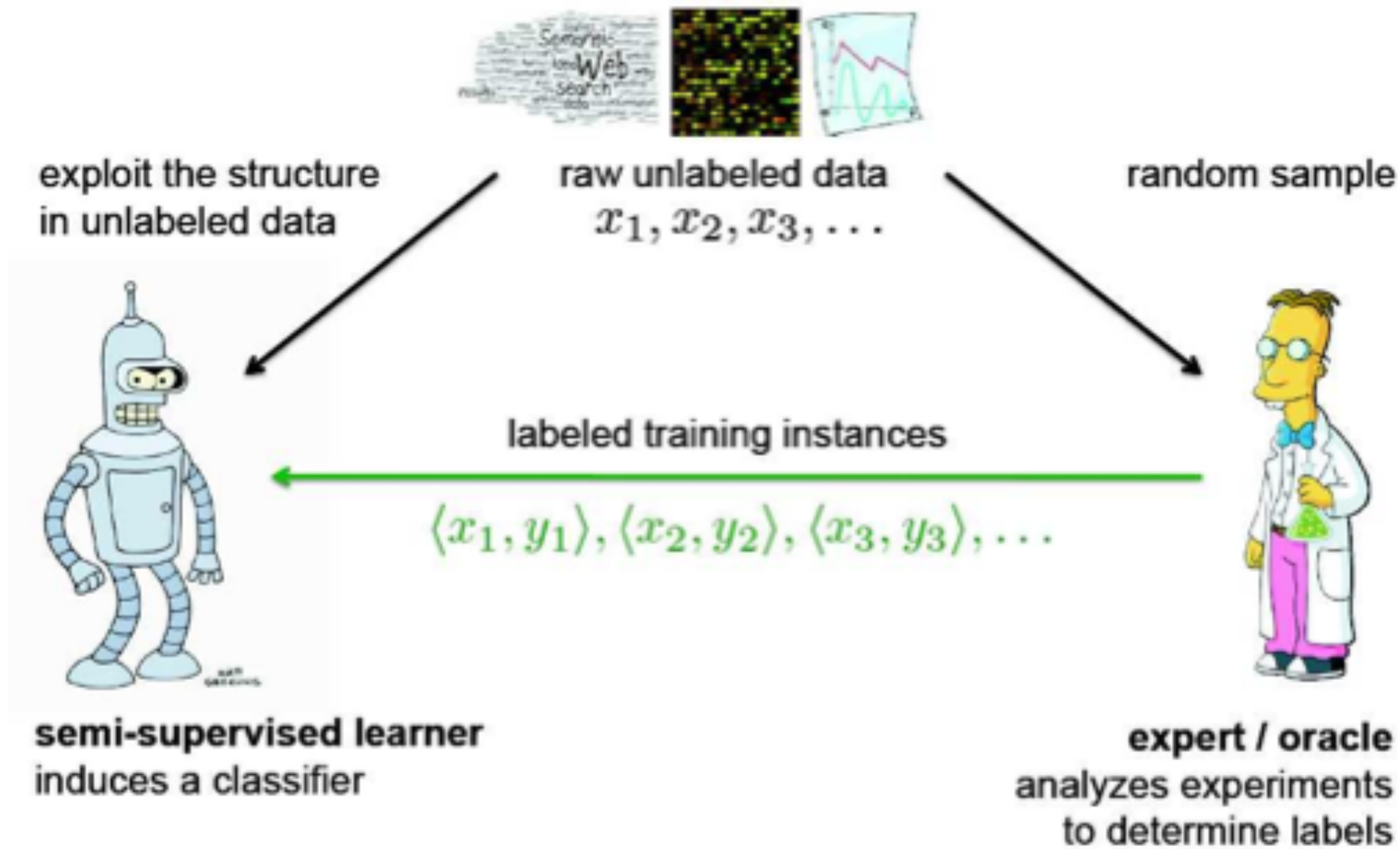
Slide credit: Piyush Rai

Unsupervised learning



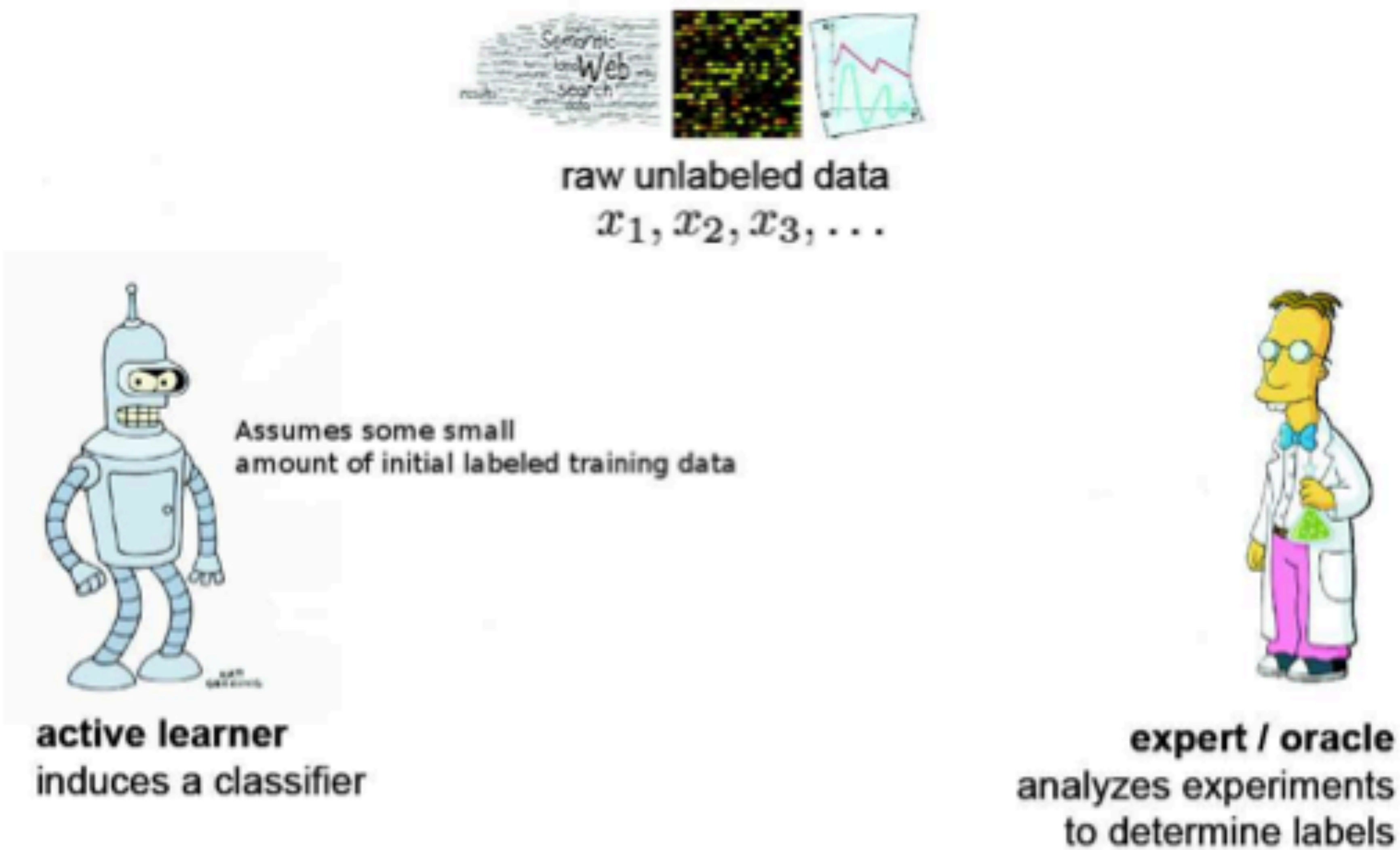
Slide credit: Piyush Rai

Semi-supervised learning



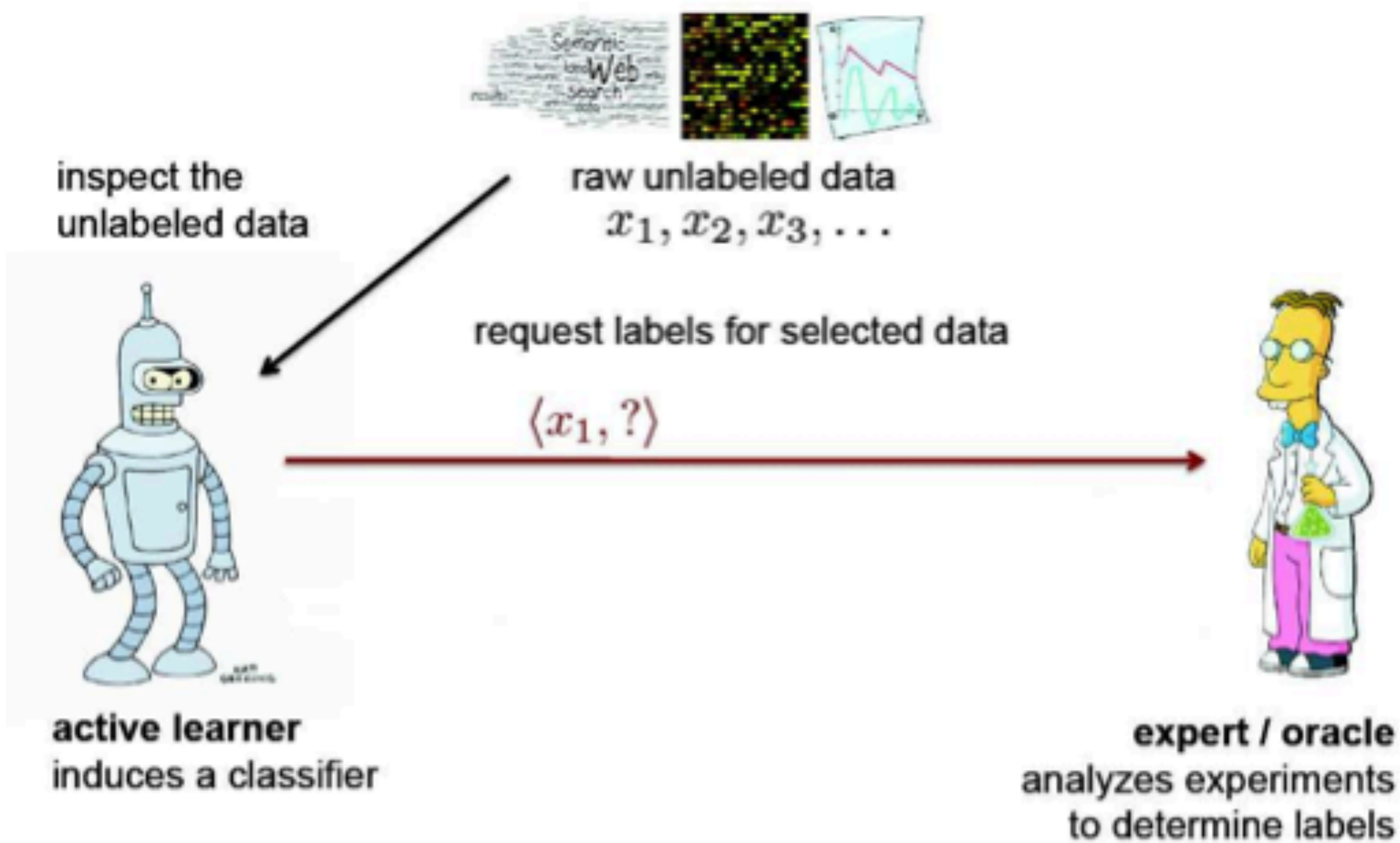
Slide credit: Piyush Rai

Active learning



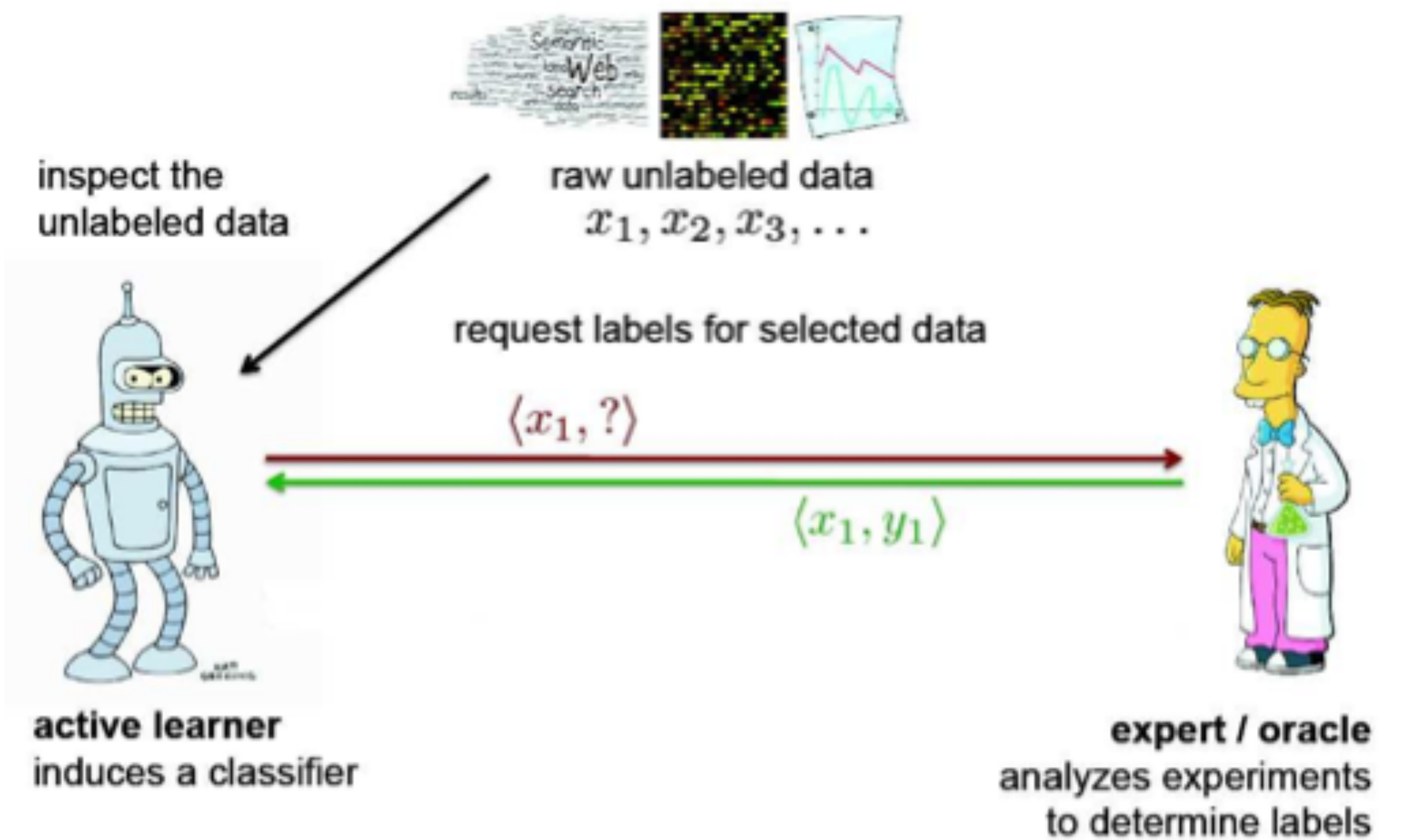
Slide credit: Piyush Rai

Active learning



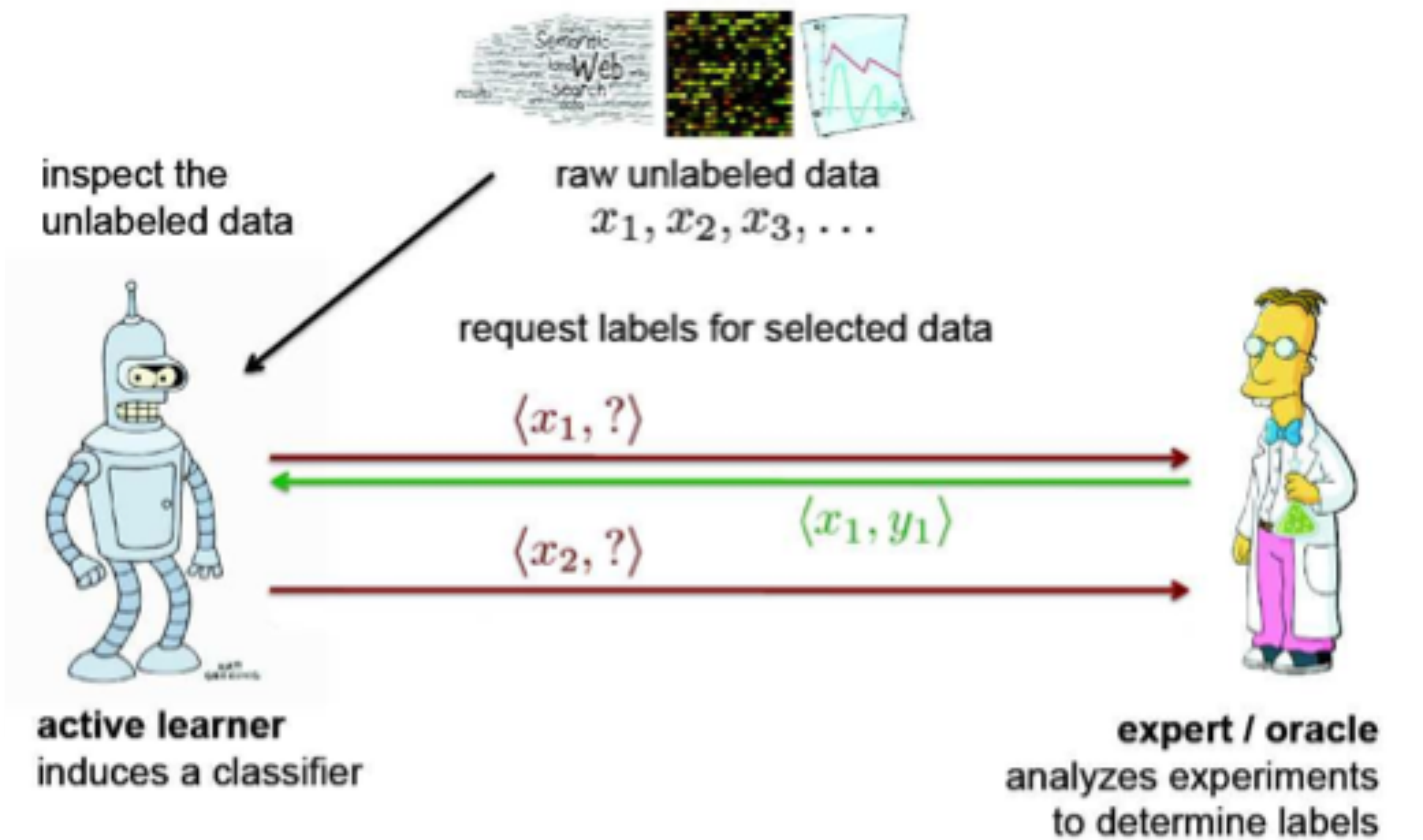
Slide credit: Piyush Rai

Active learning



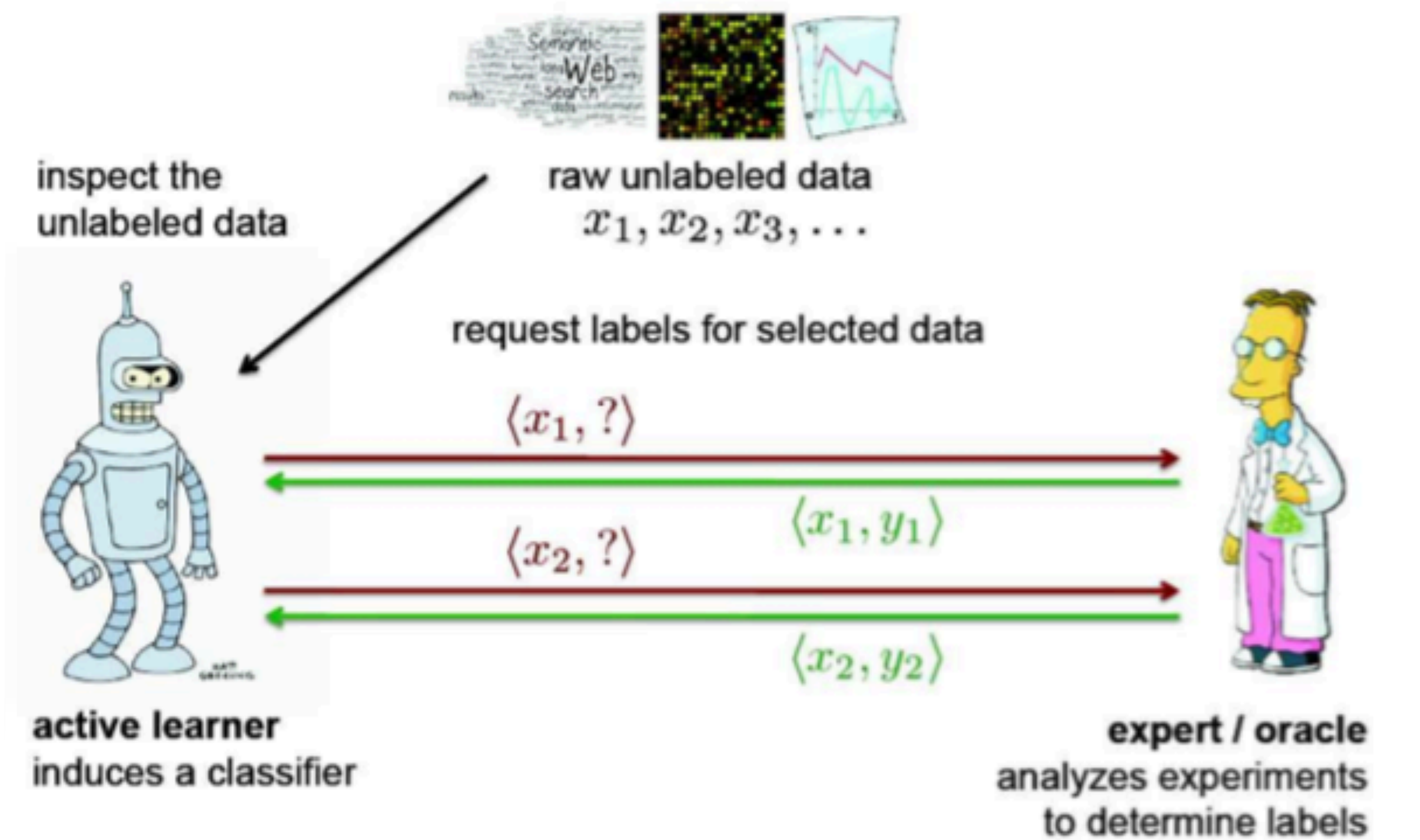
Slide credit: Piyush Rai

Active learning



Slide credit: Piyush Rai

Active learning

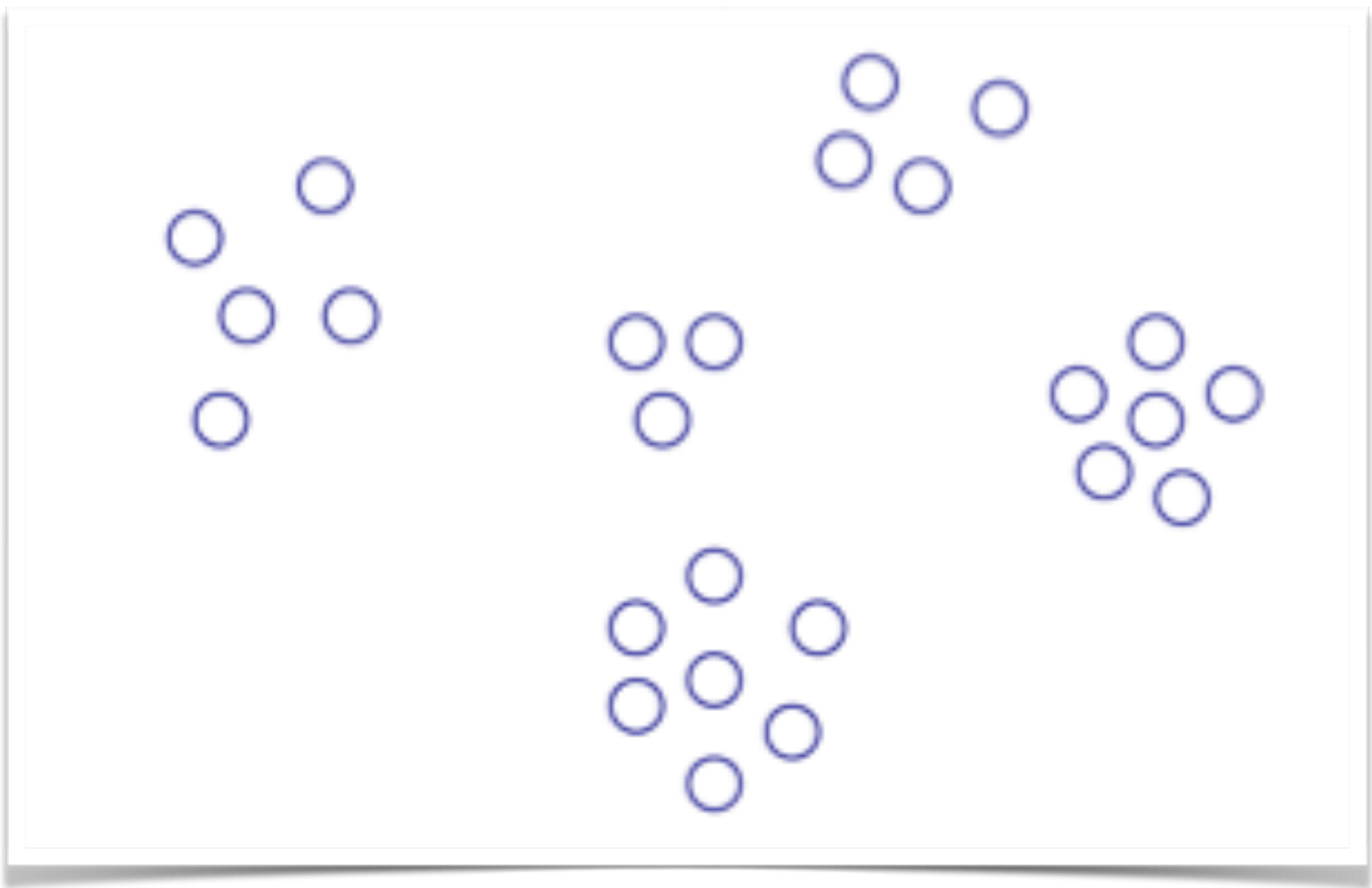


Slide credit: Piyush Rai

Motivation

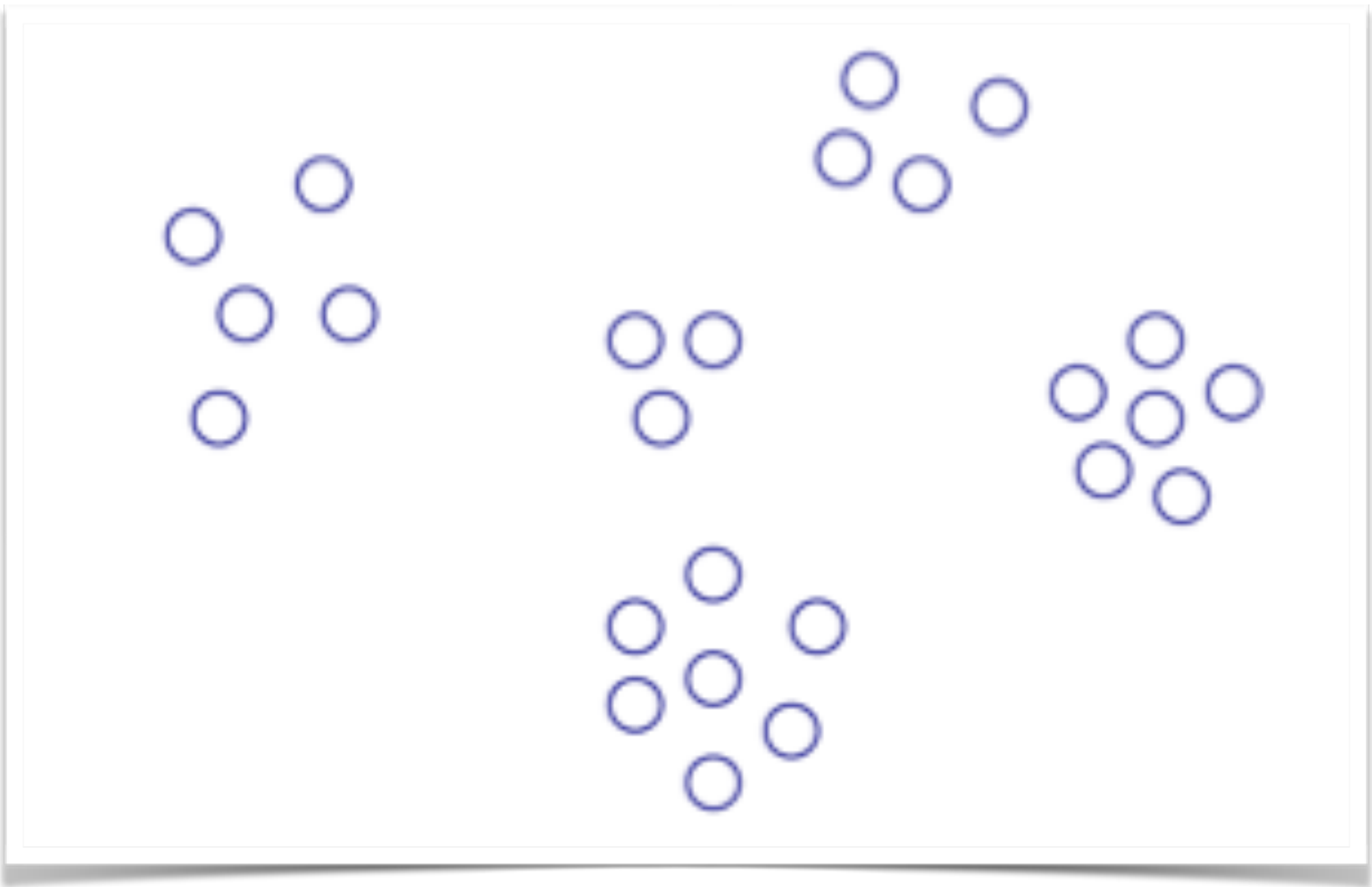
- Labels are expensive
- Maybe we can reduce the cost of training a good model by picking training examples **cleverly**

Why active learning?



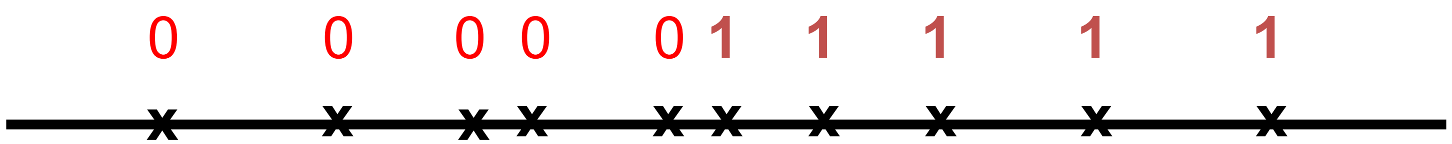
Suppose classes looked like this

Why active learning?



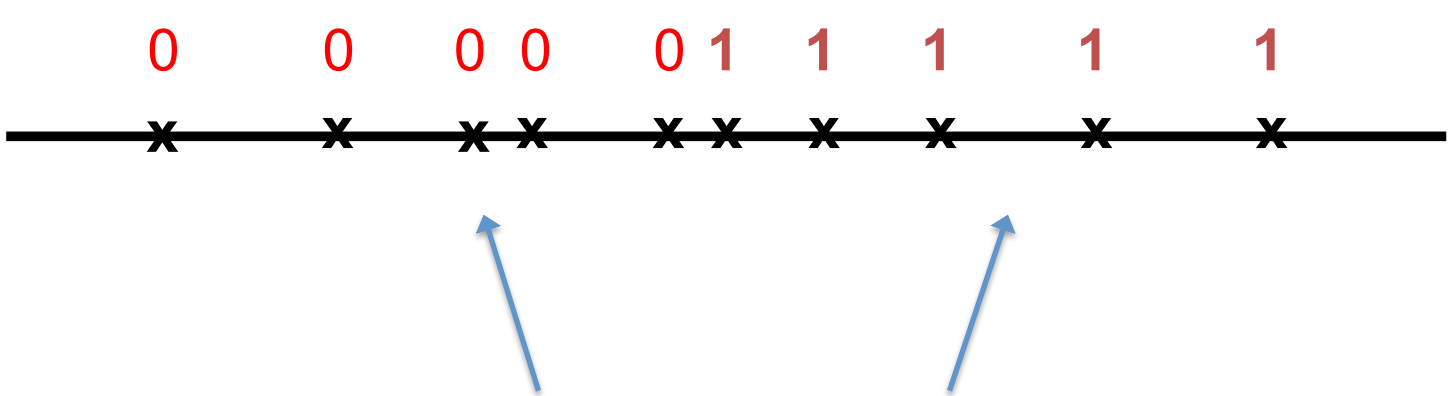
Suppose classes looked like this
We only need 5 labels!

Why active learning?



Example from Daniel Ting

Why active learning?

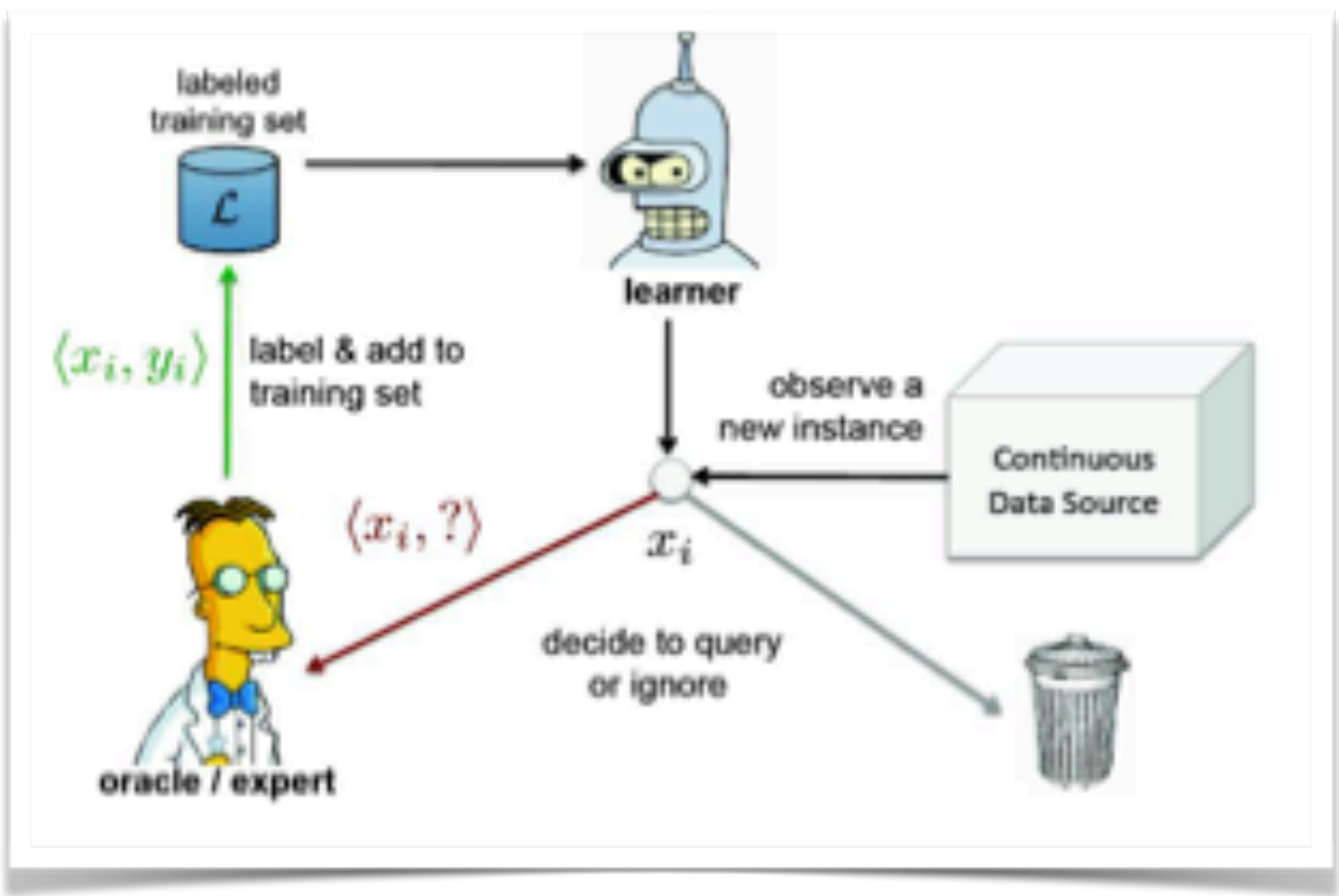


Labeling points out here is not helpful!

Example from Daniel Ting

Types of AL

- **Stream-based active learning** Consider one unlabeled instance at a time; decide whether to query for its label (or to ignore it).

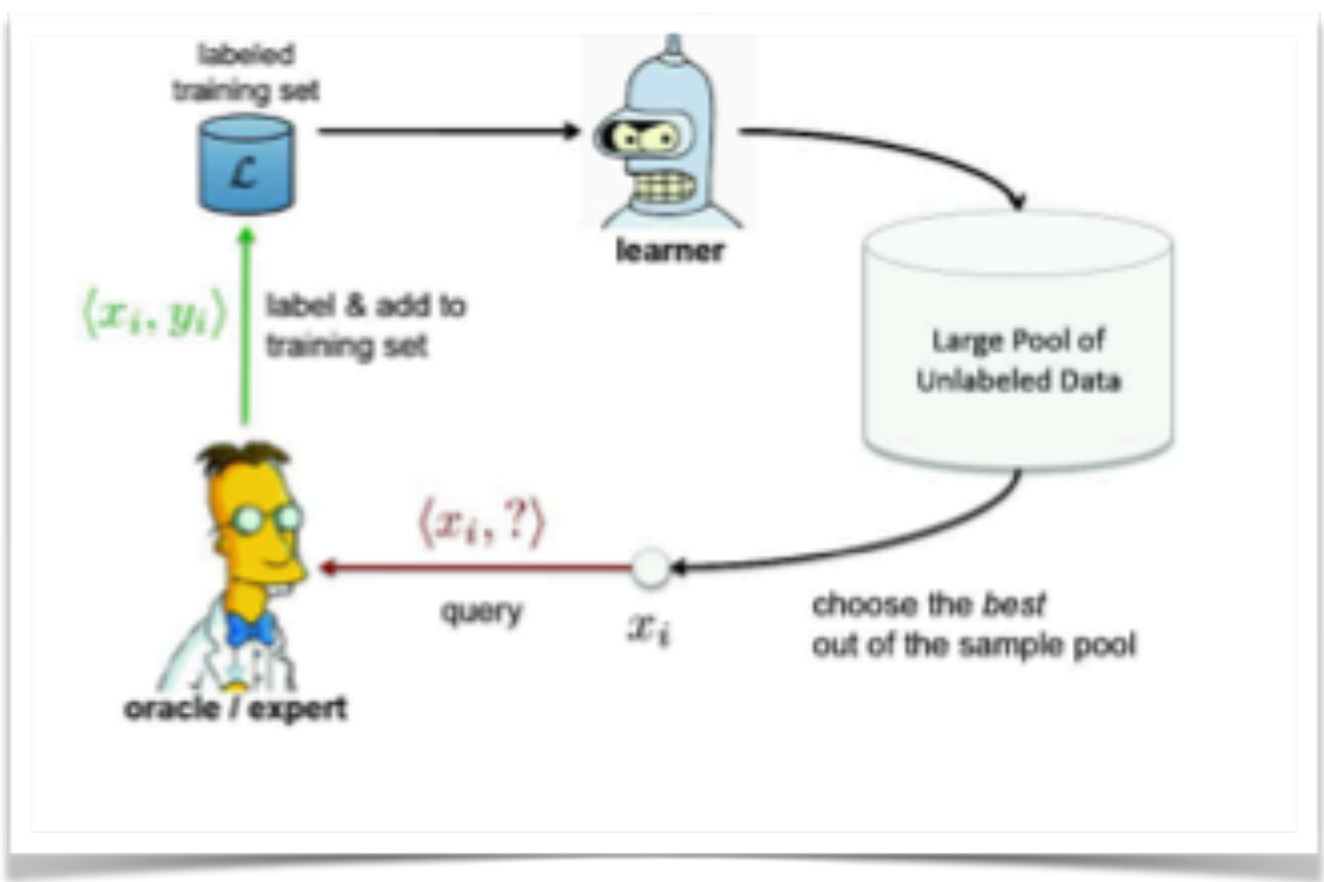


Types of AL

- **Pool-based active learning** Given a large “pool” of unlabeled examples, rank these with some heuristic that aims to capture informativeness

Types of AL

- **Pool-based active learning** Given a large "pool" of unlabeled examples, rank these with some heuristic that aims to capture informativeness



Pool based AL

- Pool-based active learning proceeds in rounds
  - Each round is associated with a current model that is learned using the labeled data seen thus far

Pool based AL

- Pool-based active learning proceeds in rounds
  - Each round is associated with a current model that is learned using the labeled data seen thus far

- The model selects the most informative example(s) remaining to be labeled at each step
  - We then pay to acquire these labels

Pool based AL

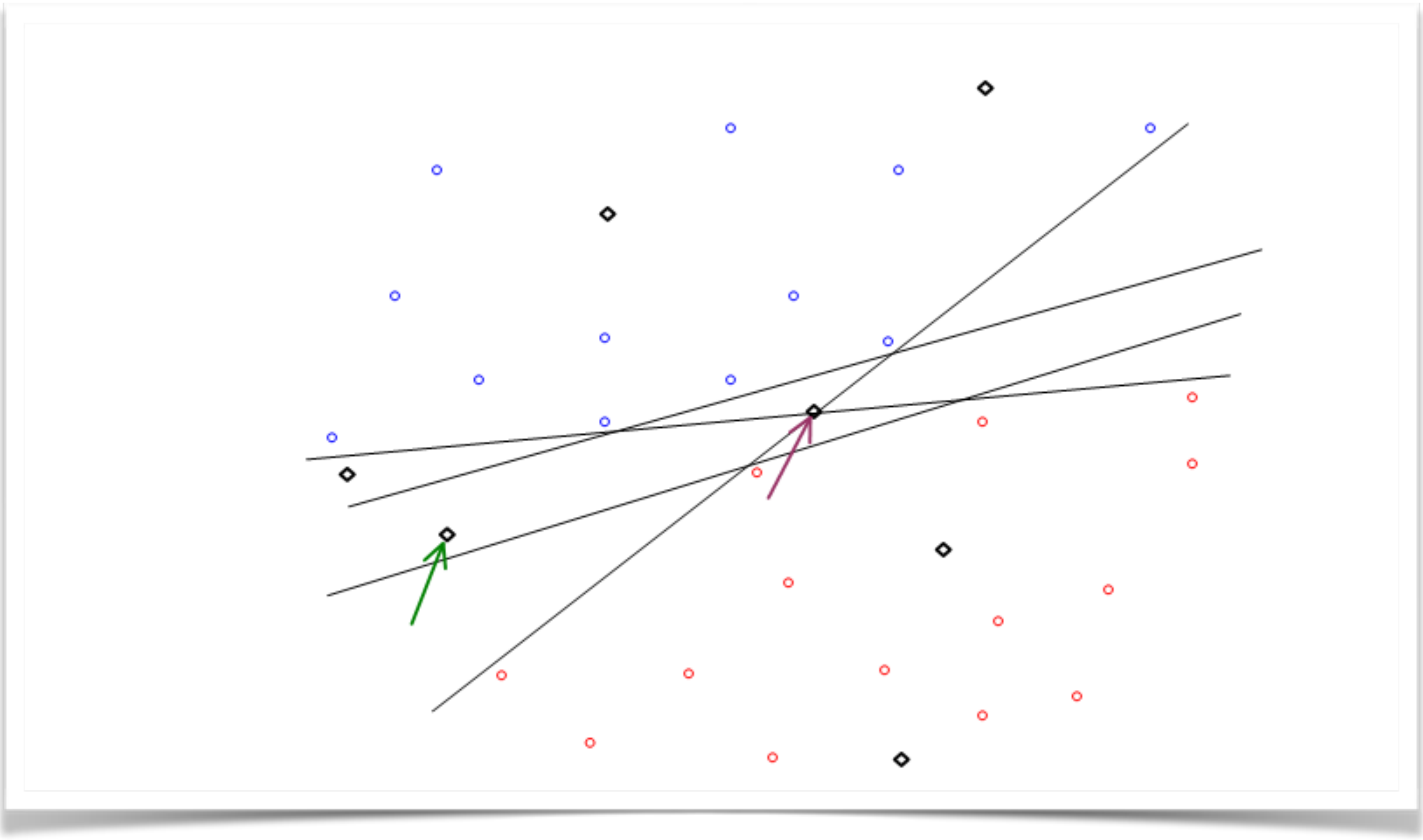
- Pool-based active learning proceeds in rounds
  - Each round is associated with a current model that is learned using the labeled data seen thus far
- The model selects the most informative example(s) remaining to be labeled at each step
  - We then pay to acquire these labels
- New labels are added to the labeled data; the model is re-trained

Pool based AL

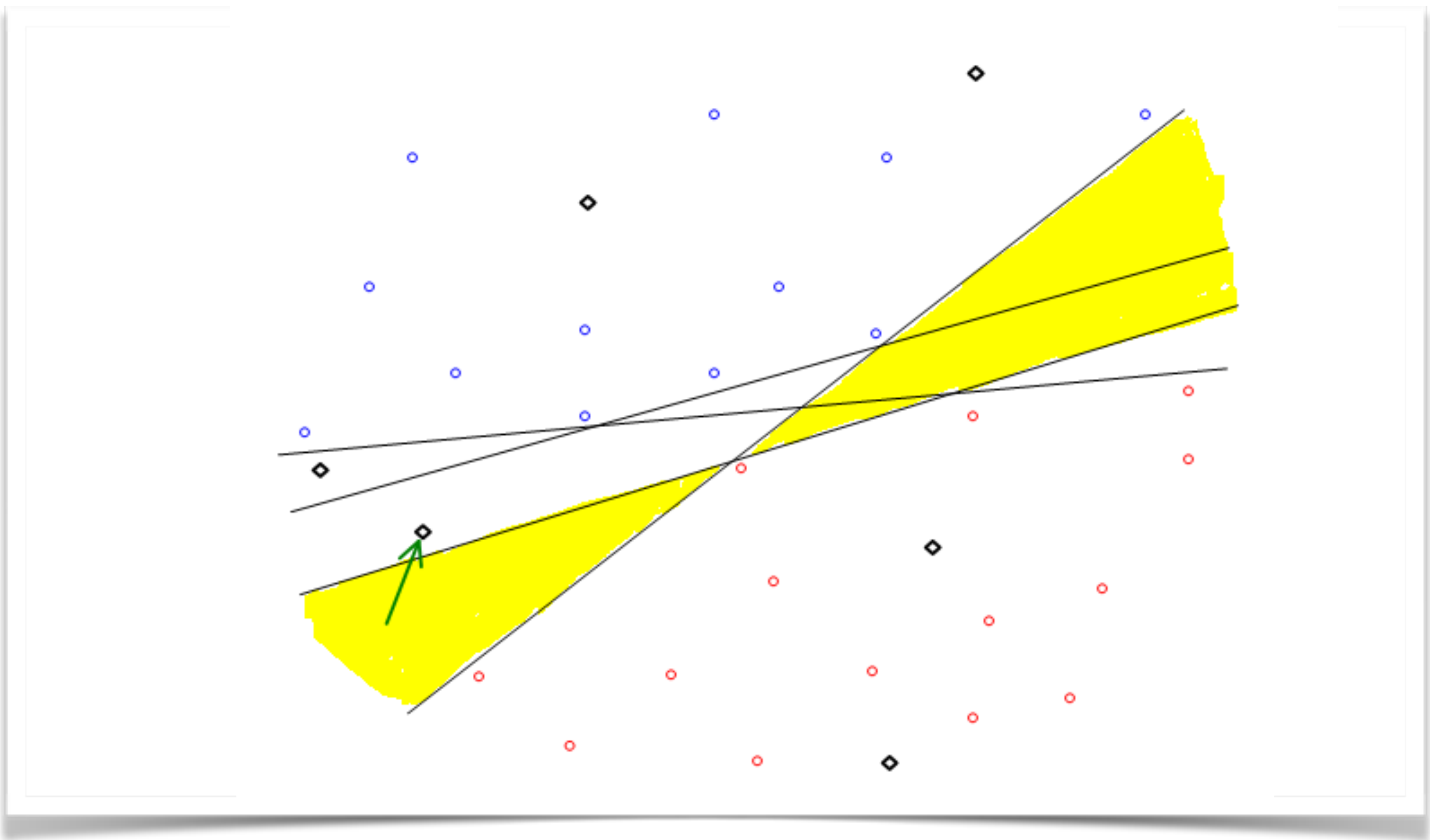
- Pool-based active learning proceeds in rounds
  - Each round is associated with a current model that is learned using the labeled data seen thus far
- The model selects the most informative example(s) remaining to be labeled at each step
  - We then pay to acquire these labels
- New labels are added to the labeled data; the model is re-trained
- We repeat this process until we are out of $$$

How might we pick ‘good’ unlabeled examples?

Query by Committee (QBC)

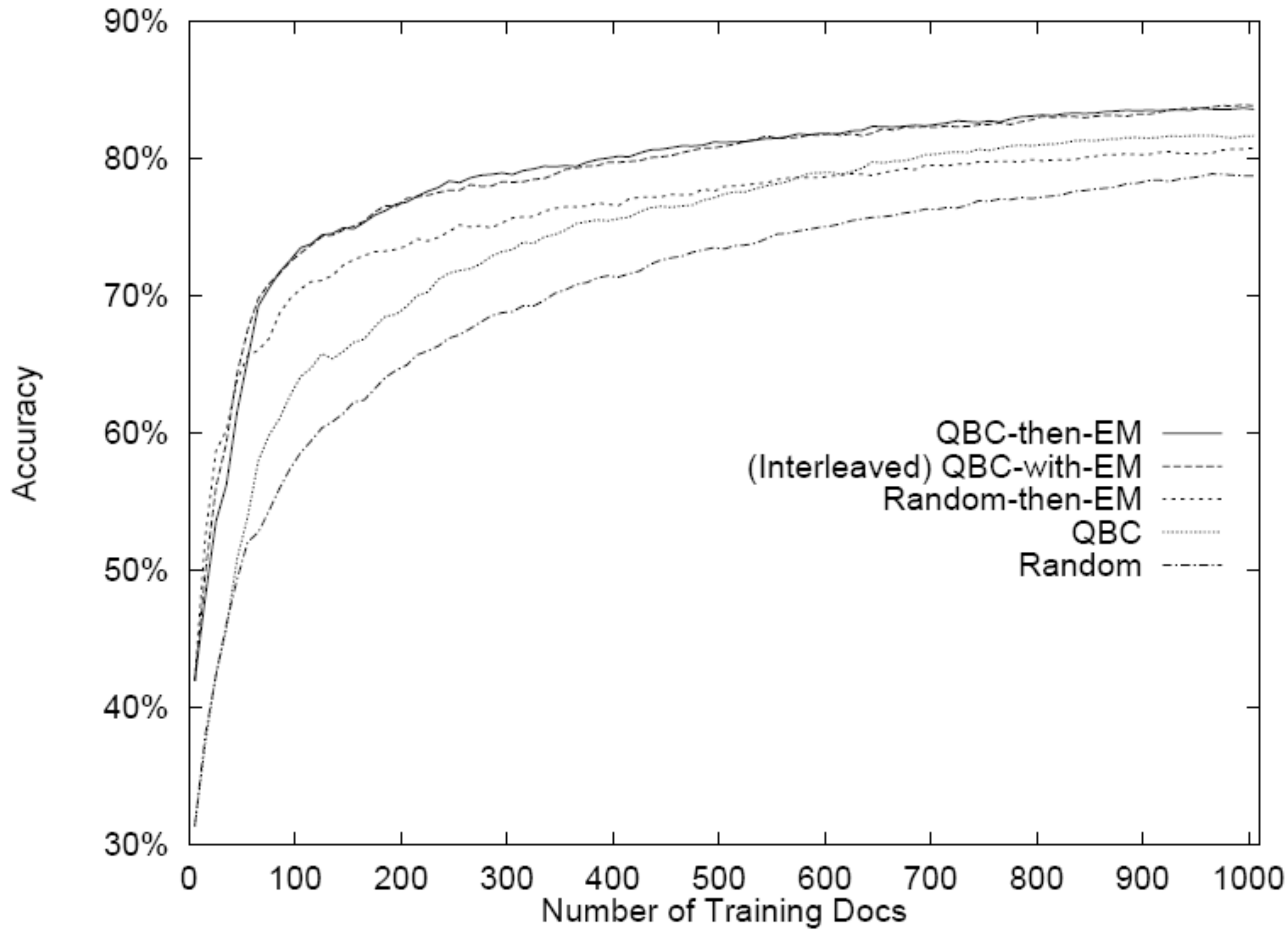


Query by Committee (QBC)



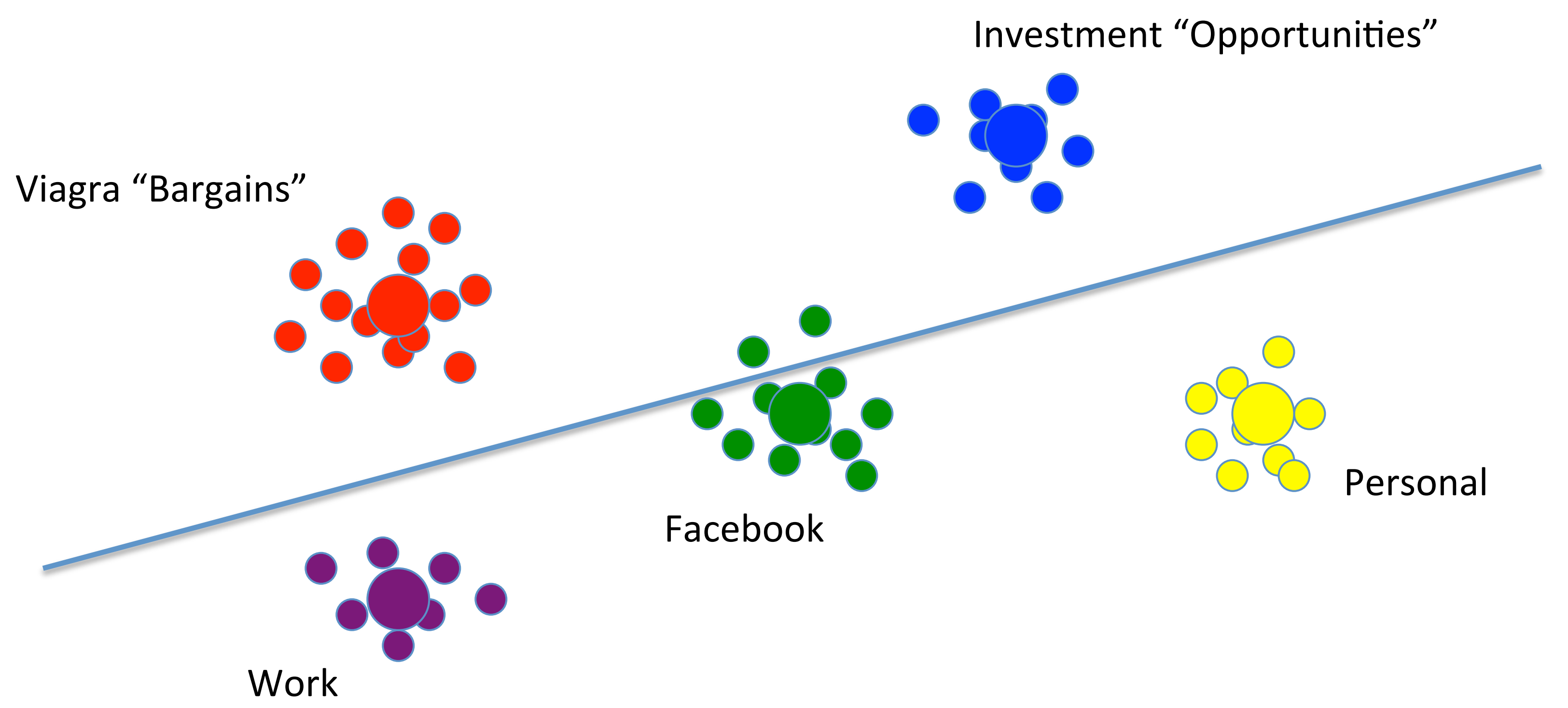
Picking point about which there is most disagreement

Query by Committee (QBC)



[McCallum & Nigam, 1998]

Pre-Clustering



If data clusters, we only require a few representative instances from each cluster to label data

[Ngyuen & Smeulders 04]

Uncertainty sampling

- Query the event that the current classifier is most **uncertain** about

Uncertainty sampling

- Query the event that the current classifier is most **uncertain** about

- Needs measure of uncertainty, probabilistic model for prediction!

Uncertainty sampling

- Query the event that the current classifier is most **uncertain** about
- Needs measure of uncertainty, probabilistic model for prediction!
- Examples:
  – Entropy
  – Least confident predicted label
  – Euclidean distance (e.g. point closest to margin in SVM)

Uncertainty sampling

$$x^* = \arg\min_x P(\hat{y}|x,\theta) = \arg\min_x \max_y P(y|x,\theta)$$

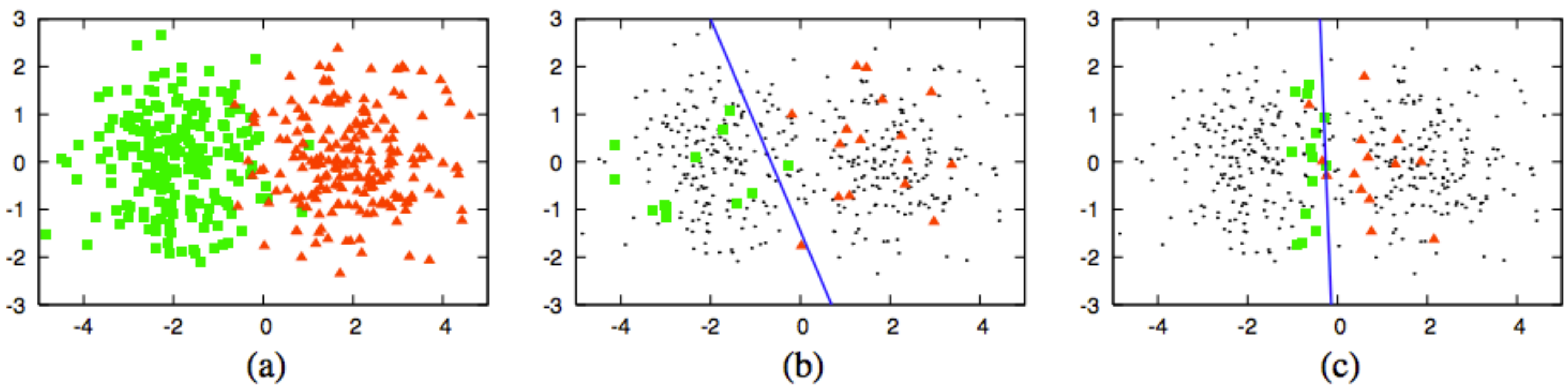


Figure 2: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain. The line represents the decision boundary of the classifier (70% accuracy). (c) A logistic regression model trained with 30 actively queried instances using uncertainty sampling (90%).

Let’s implement this…

(“in class” exercise on *active learning*)

In class exercise 3/22

Availability: Item is hidden from students. It will be available after Mar 24, 2020 8:00 AM.

Start: https://colab.research.google.com/drive/19cAl2TQ-CBEG_GJjg-Hc-xuO6Drs6PCm

Practical Obstacles to Deploying Active Learning

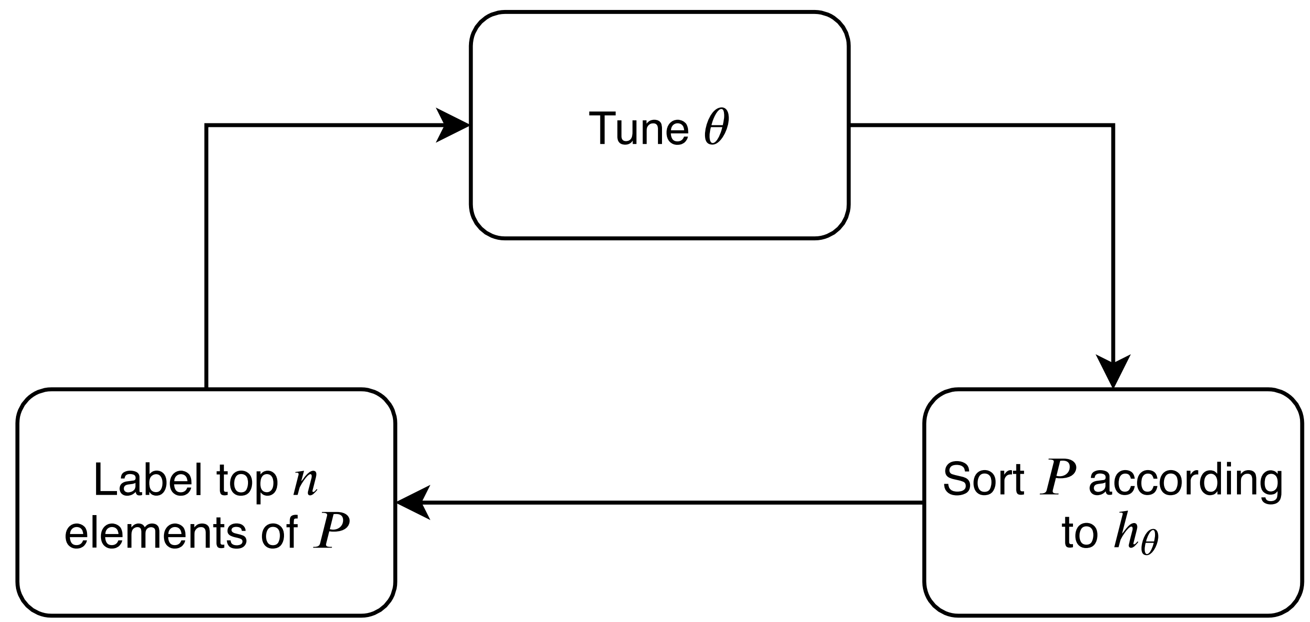
David Lowell
Northeastern University

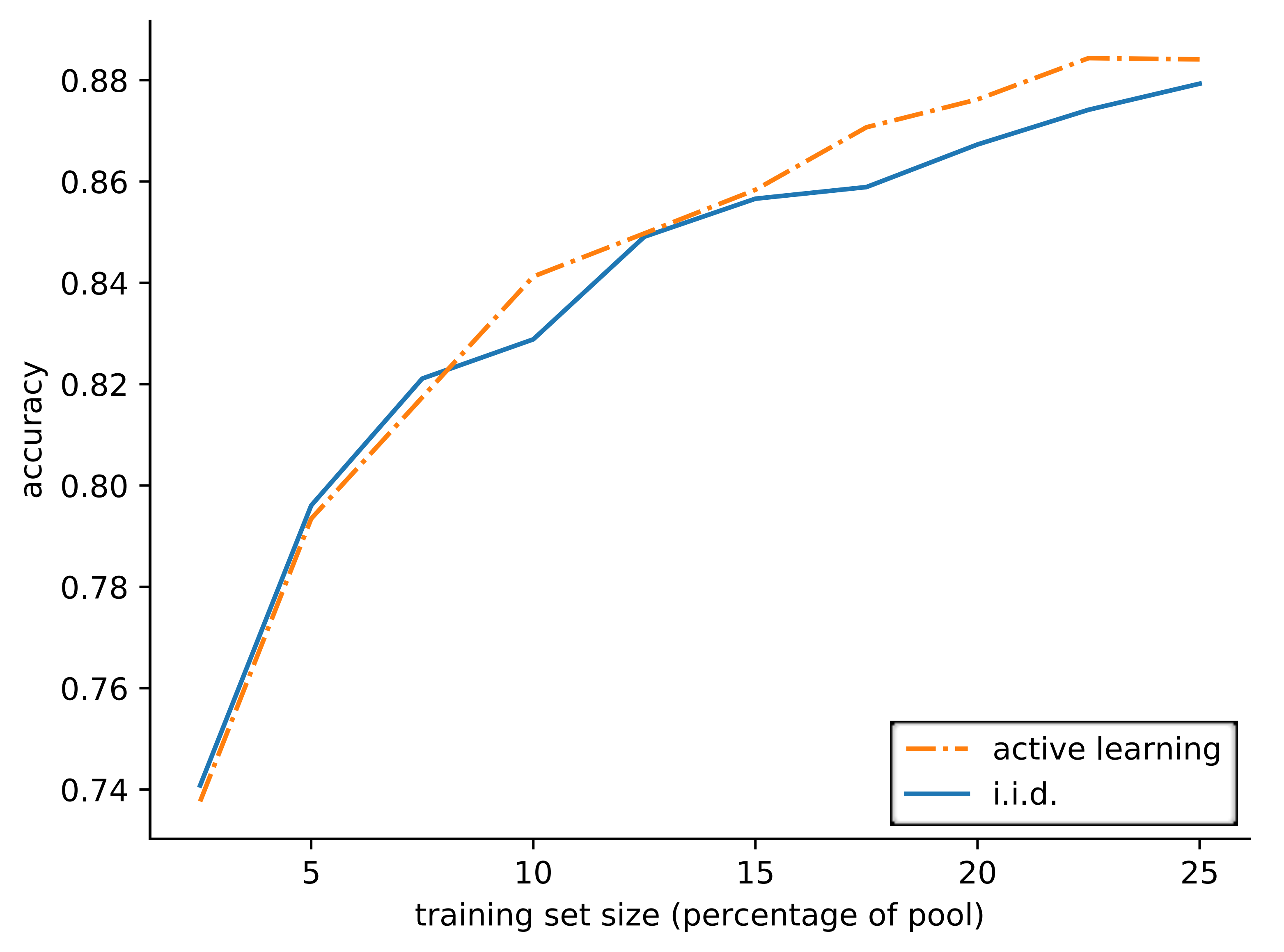
Zachary C. Lipton
Carnegie Mellon University

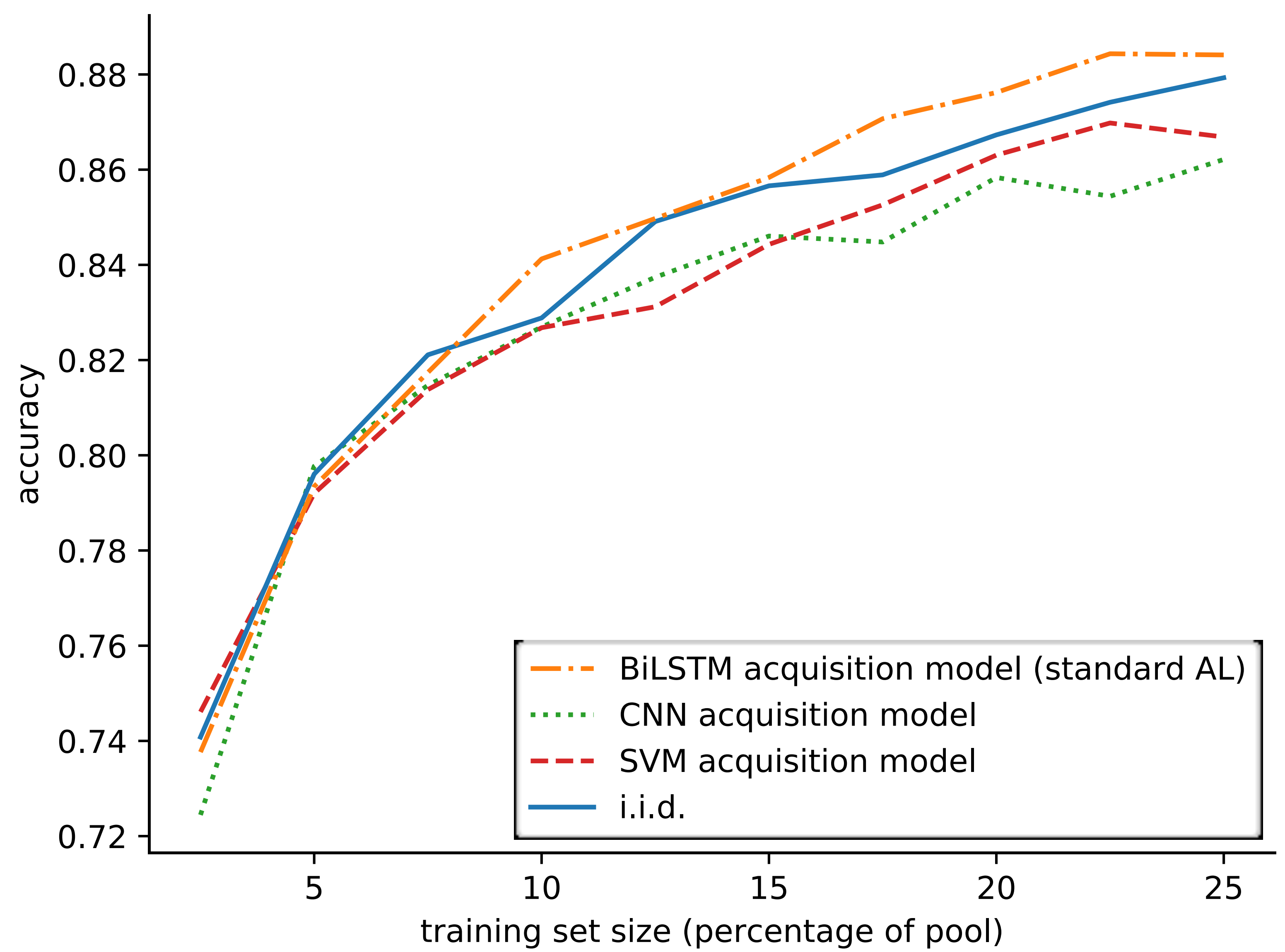
Byron C. Wallace
Northeastern University

Given

- Pool of unlabeled data *P*
- Model parameterized by θ
- A sorting heuristic *h*




accuracy
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
5
10
15
20
25
training set size (percentage of pool)
active learning
i.i.d.



accuracy
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
5
10
15
20
25
training set size (percentage of pool)
BiLSTM acquisition model (standard AL)
CNN acquisition model
SVM acquisition model
i.i.d.

Some issues

- Users must *choose* a single heuristic (AL strategy) from many choices before acquiring more data

- Active learning *couples* datasets to the model used at acquisition time

Experiments

Active Learning involves:

- A data pool
- An acquisition model and function
- A “successor” model (to be trained)

Tasks & datasets

Classification

Movie reviews, Subjectivity/objectivity, Customer reviews, Question type classification

Sequence labeling (NER)

CoNLL, OntoNotes

Models

Classification

SVM, CNN, BiLSTM

Sequence labeling (NER)

CRF, BiLSTM-CNN

Uncertainty sampling

$$\underset{\mathbf{x} \in \mathcal{U}}{\operatorname{argmax}} - \sum_j P(y_j|\mathbf{x}) \log P(y_j|\mathbf{x})$$

(For sequences)

$$\max_{y_1,\ldots,y_n} \frac{1}{n} \sum_{i=1}^{n} \log P(y_i | y_1, \ldots, y_{n-1}, \mathbf{x})$$

Query By Committee (QBC)

$$\underset{\mathbf{x}\in\mathcal{U}}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} \sum_{j} P_c(y_j|\mathbf{x}) \log \frac{P_c(y_j|\mathbf{x})}{P_C(y_j|\mathbf{x})}$$

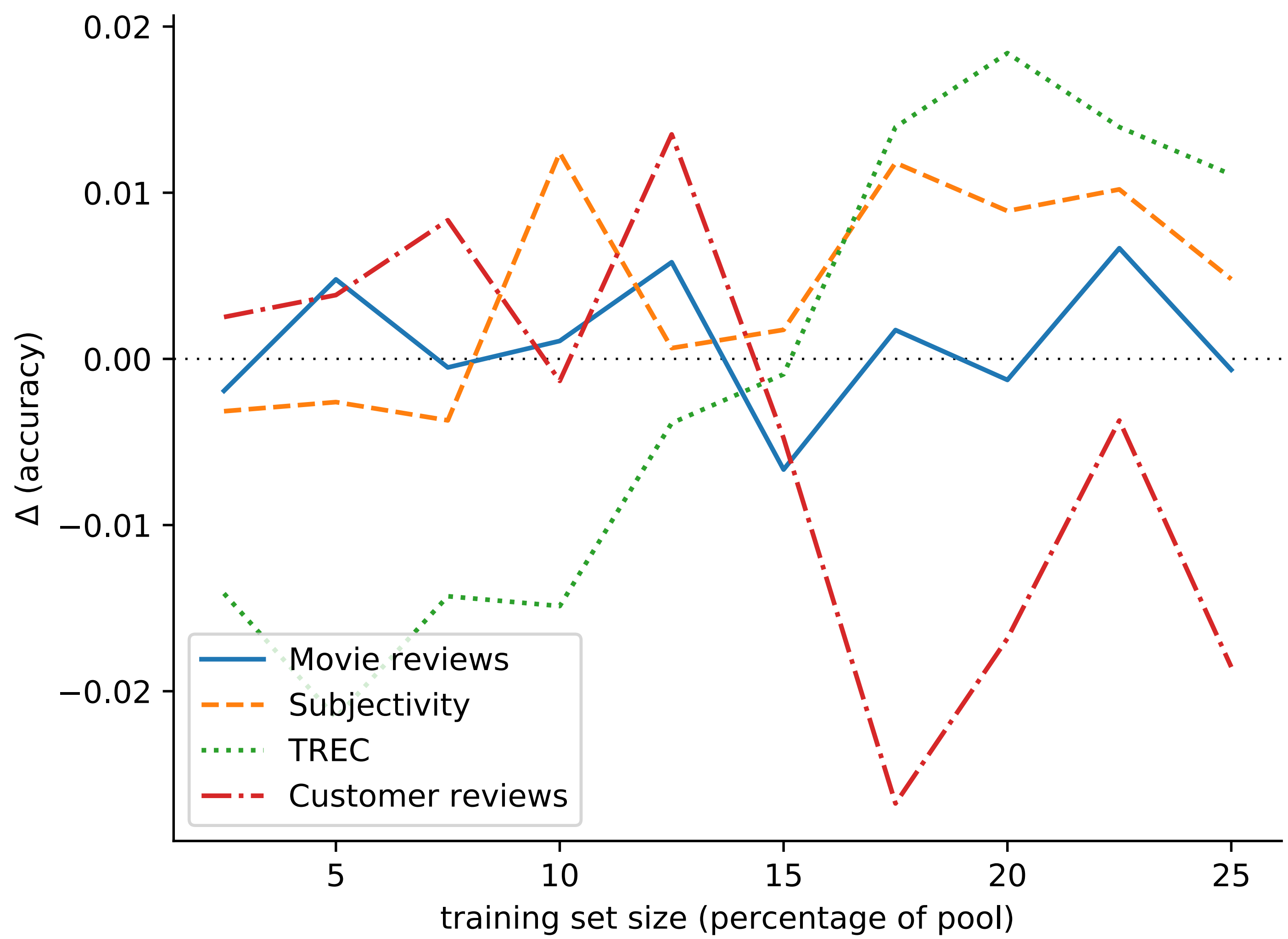
(For sequences)

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{m}\frac{V(y_i,m)}{C}\log\frac{V(y_i,m)}{C}$$

Results

- 75.0%: there exists a heuristic that outperforms i.i.d.
- 60.9%: a specific heuristic outperforms i.i.d.
- 37.5%: transfer of actively acquired data outperforms i.i.d.

- But, active learning consistently outperforms i.i.d. for sequential tasks



(a) Performance of AL relative to i.i.d. across corpora.

Results

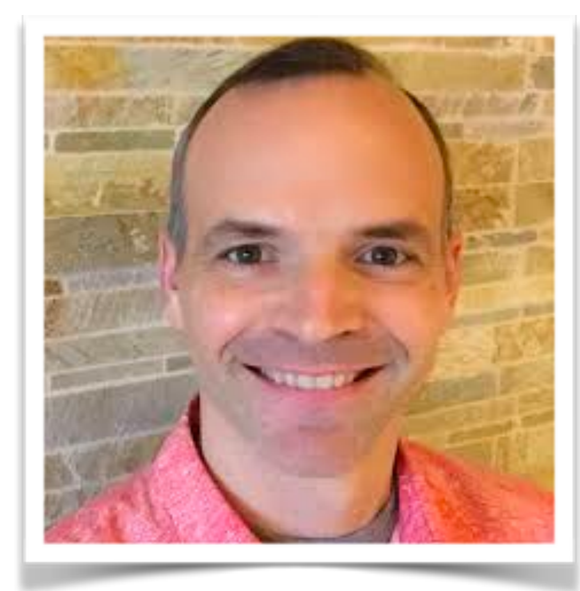
It is difficult to characterize when AL will be successful

Trends:

- Uncertainty with SVM or CNN
- BALD with CNN
- AL transfer leads to poor results

Crowdsourcing

slides derived from *Matt Lease*



Crowdsourcing

- In ML, *supervised learning* still dominates (despite the various innovations in self-/un-supervised learning we have seen in this class

Crowdsourcing

- In ML, *supervised learning* still dominates (despite the various innovations in self-/un-supervised learning we have seen in this class
- Supervision is expensive; modern (deep) models need lots of it

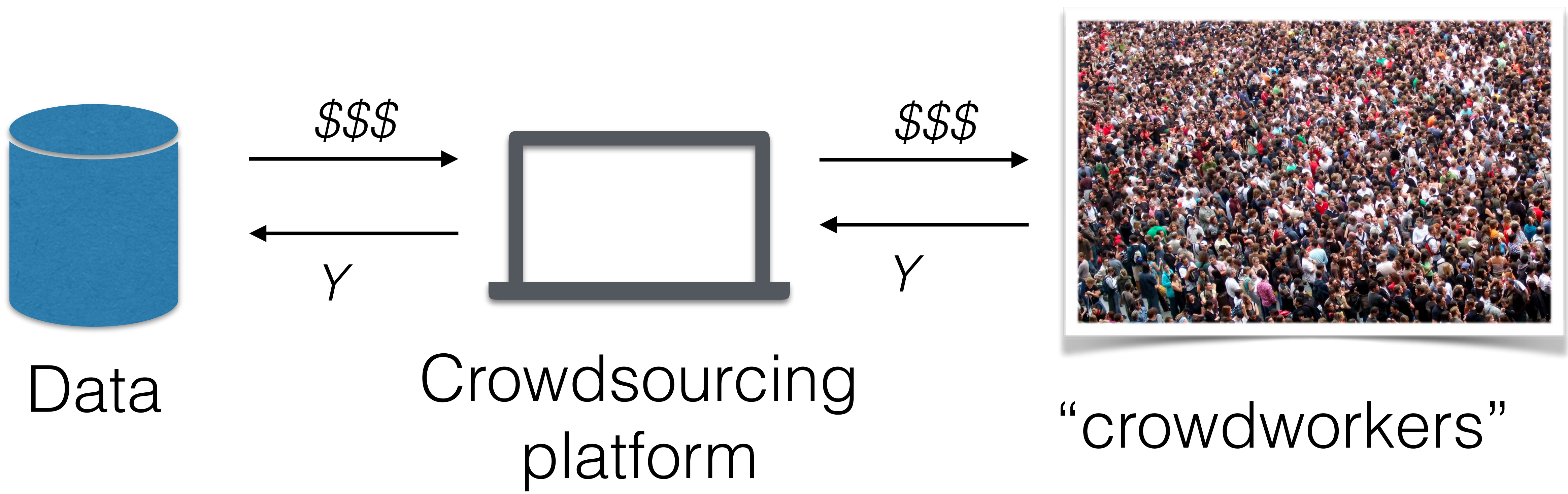
Crowdsourcing

- In ML, *supervised learning* still dominates (despite the various innovations in self-/un-supervised learning we have seen in this class

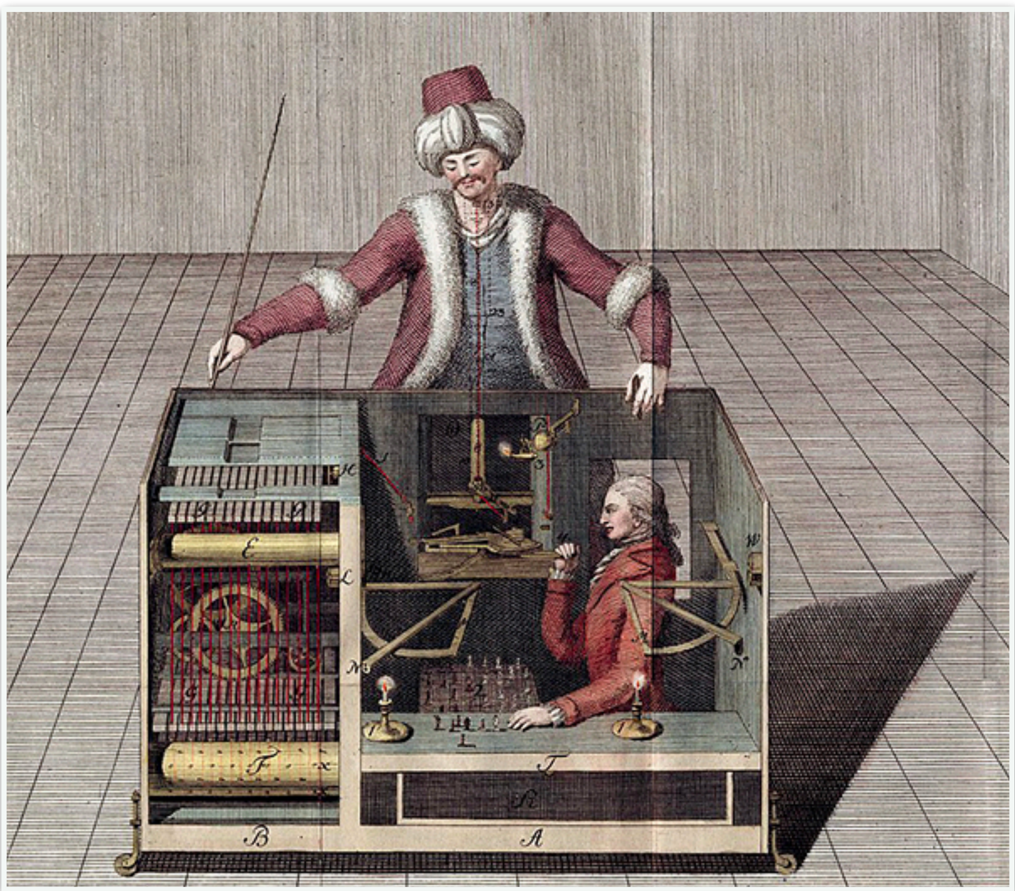
- Supervision is expensive; modern (deep) models need lots of it

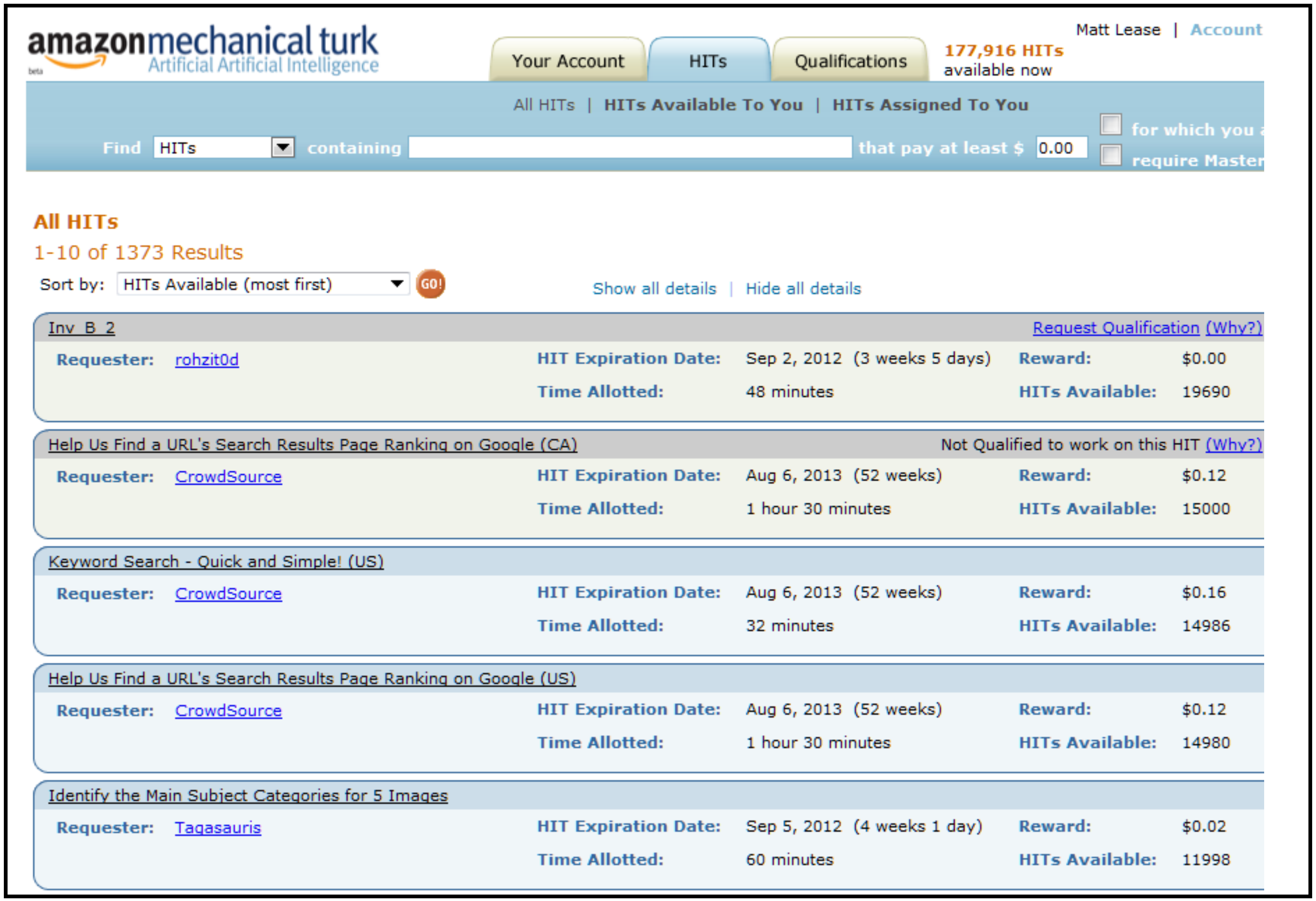
- One use of **crowdsourcing** is collecting lots of annotations, on the cheap

Crowdsourcing



Crowdsourcing




amazonmechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
177,916 HITs
available now
Matt Lease | Account
All HITs | HITs Available To You | HITs Assigned To You
Find HITs containing that pay at least $ 0.00
for which you a
require Master
All HITs
1-10 of 1373 Results
Sort by: HITs Available (most first) GO!
Show all details | Hide all details
Inv_B_2
Request Qualification (Why?)
Requester: rohzit0d
HIT Expiration Date: Sep 2, 2012 (3 weeks 5 days)
Reward: $0.00
Time Allotted: 48 minutes
HITs Available: 19690
Help Us Find a URL's Search Results Page Ranking on Google (CA)
Not Qualified to work on this HIT (Why?)
Requester: CrowdSource
HIT Expiration Date: Aug 6, 2013 (52 weeks)
Reward: $0.12
Time Allotted: 1 hour 30 minutes
HITs Available: 15000
Keyword Search - Quick and Simple! (US)
Requester: CrowdSource
HIT Expiration Date: Aug 6, 2013 (52 weeks)
Reward: $0.16
Time Allotted: 32 minutes
HITs Available: 14986
Help Us Find a URL's Search Results Page Ranking on Google (US)
Requester: CrowdSource
HIT Expiration Date: Aug 6, 2013 (52 weeks)
Reward: $0.12
Time Allotted: 1 hour 30 minutes
HITs Available: 14980
Identify the Main Subject Categories for 5 Images
Requester: Tagasauris
HIT Expiration Date: Sep 5, 2012 (4 weeks 1 day)
Reward: $0.02
Time Allotted: 60 minutes
HITs Available: 11998

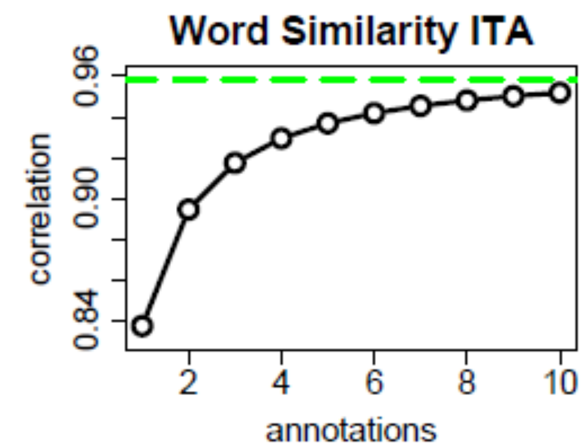
Cheap and Fast — But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks

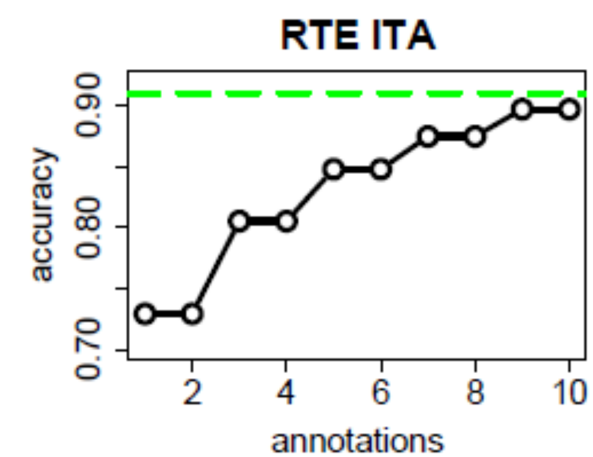
Rion Snow[†] **Brendan O'Connor**[‡] **Daniel Jurafsky**[§] **Andrew Y. Ng**[†]

[†]Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion,ang}@cs.stanford.edu

[‡]Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

[§]Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu





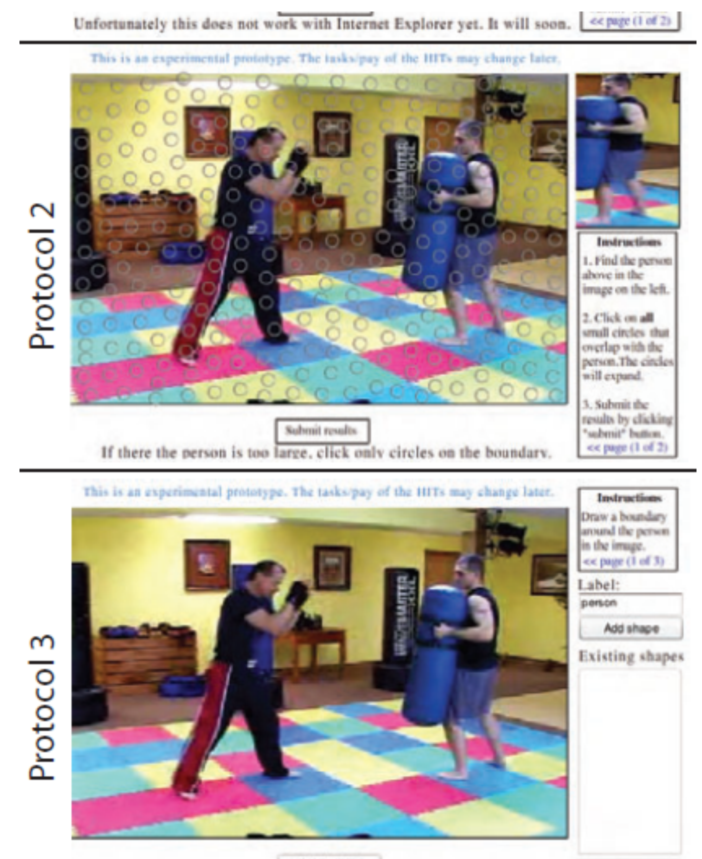
Our evaluation of non-expert labeler data vs. expert annotations for five tasks found that for many tasks only a small number of non- expert annotations per item are necessary to equal the performance of an expert annotator.

Computer Vision:
Sorokin & Forsythe (CVPR 2008)

- 4K labels for US $60

Exp	Task	img	labels	cost USD	time	effective pay/hr
1	1	170	510	$8	750m	$0.76
2	2	170	510	$8	380m	$0.77
3	3	305	915	$14	950m	$0.41[1]
4	4	305	915	$14	150m	$1.07
5	4	337	1011	$15	170m	$0.9
Total:		982	3861	$59		

Table 1. **Collected data.** In our five experiments we have collected **3861** labels for 982 distinct images for only **US $59**. In experiments 4 and 5 the throughput exceeds 300 annotations per hour even at low ($1/hour) hourly rate. We expect further increase in throughput as we increase the pay to effective market rate.



Dealing with noise

Problem Crowd annotations are often noisy

Dealing with noise

Problem Crowd annotations are often noisy

One way to address: collect independent annotations from multiple workers

Dealing with noise

Problem Crowd annotations are often noisy

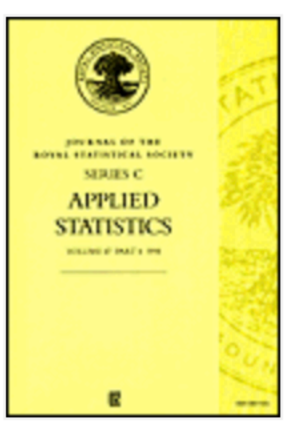
One way to address: collect independent annotations from multiple workers

But then how to combine these?

Dawid-Skene

Define a simple probabilistic model of worker annotations, conditioned on latent “true” labels for instances

Can easily estimate via Expectation-Maximization



JOURNAL ARTICLE

Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm

A. P. Dawid and A. M. Skene

Journal of the Royal Statistical Society. Series C (Applied Statistics)

Vol. 28, No. 1 (1979), pp. 20-28

I instances

J labelers

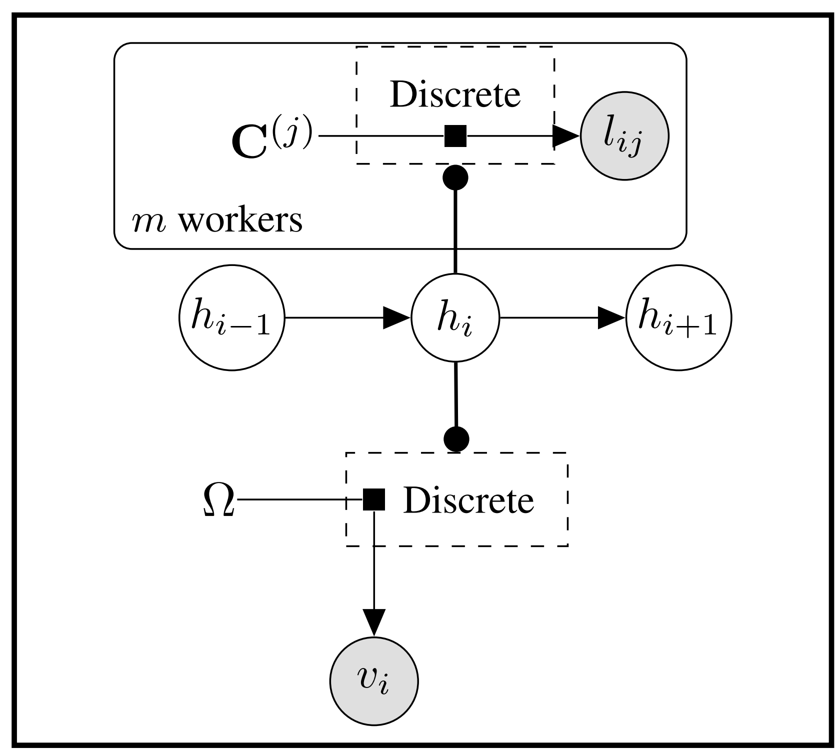
$$p(y|\theta,\pi) \;=\; \prod_{i=1}^{I} \sum_{k=1}^{K} \left(\mathsf{Categorical}(z_i|\pi) \prod_{j=1}^{J} \mathsf{Categorical}(y_{i,j}|\theta_{j,z[i]}) \right)$$

K categories (classes)

Aggregating and Predicting Sequence Labels from Crowd Annotations

An T. Nguyen[1] **Byron C. Wallace**[2] **Junyi Jessy Li**[3] **Ani Nenkova**[3] **Matthew Lease** [1]

[1]University of Texas at Austin, [2]Northeastern University,
[3]University of Pennsylvania,
atn@cs.utexas.edu, byron@ccs.neu.edu,
{ljunyi|nenkova}@seas.upenn.edu, ml@utexas.edu



"Citizen Science"

Evidence-based Medicine

Become an EMBASE screener - Cochrane's innovative EMBASE project is now open for all budding volunteers!

The EMBASE project provides an opportunity for new and potential contributors to get involved with Cochrane work by diving into a task that needs doing. **No prior experience is necessary as the task supports a 'learn as you do' approach.**

The project's purpose is to identify reports of randomised controlled trials (RCTs) and quasi-RCTs from EMBASE for publication in the Cochrane Central Register of Controlled Trials (CENTRAL). It is run by a team from Metaxis Ltd, (developer of the Cochrane Register of Studies), the Cochrane Dementia and Cognitive Improvement Group, and York Health Economics Consortium (YHEC).

A crucial part of the project was to develop and implement a screening task, and the innovative bit is that this task is crowd-sourced. **A web-based screening tool has been developed so that anyone, with access to the internet, can join the collective effort to screen the search results for relevance within CENTRAL.** A quality-control system has been developed so that all records will be viewed by at least two screeners. Records viewed by 'novice' screeners will need three consecutive agreements on the record's relevance for it to then be either published in CENTRAL or 'rejected'. Disagreements will be arbitrated by experts. All new screeners have to complete a small, interactive test set of records before progressing to 'live' records.

Task routing

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

Predicting Annotation Difficulty to Improve Task Routing and Model Performance for Biomedical Information Extraction

Yinfei Yang
Google AI
yinfeiy@google.com

Oshin Agarwal
University of Pennsylvania
oagarwal@seas.upenn.edu

Chris Tar
Google AI
ctar@google.com

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Crowdsourcing takeaways

- If you're in a position of needing to acquire supervision (annotations), you'll probably want to use crowdsourcing
- Invest in good task design and think about how you will aggregate individual annotations
- It may be worth investing in a small set of "expert" annotations as well