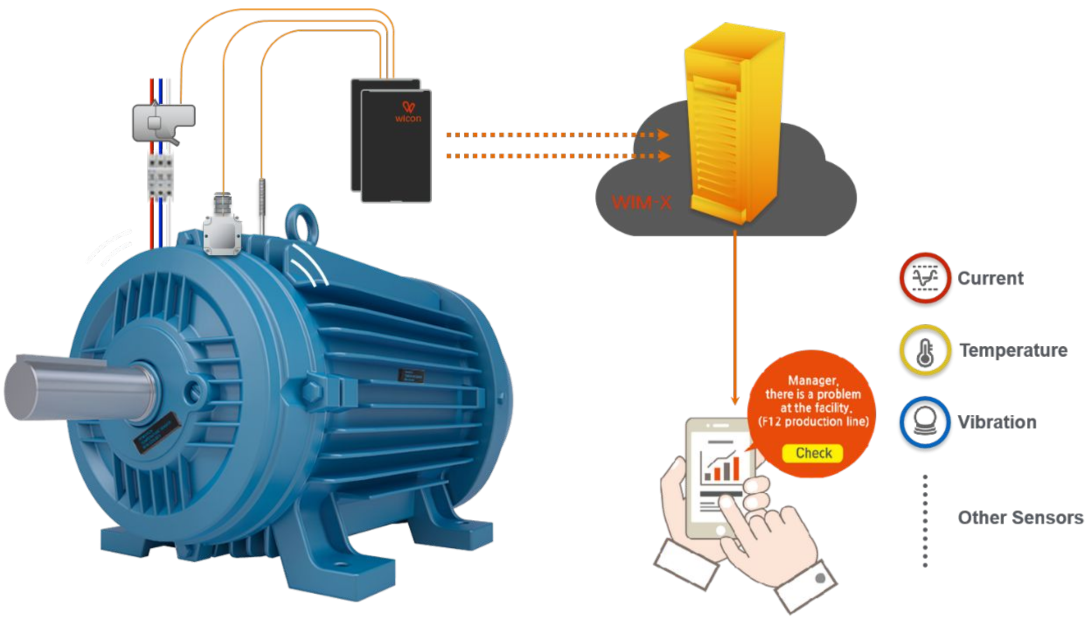
Active Learning for Sensing Applications

Modified Piyush Rai's slides from CS5350/6350: Machine Learning

Active Learning Literature Survey, Burr Settles, 2010

Example: Machine Monitoring

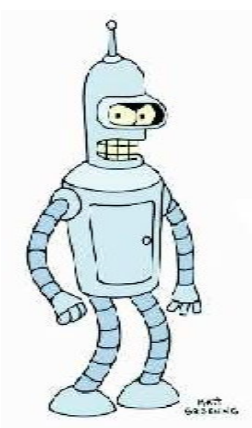
Labeling machine states by visually examining sensor data stream

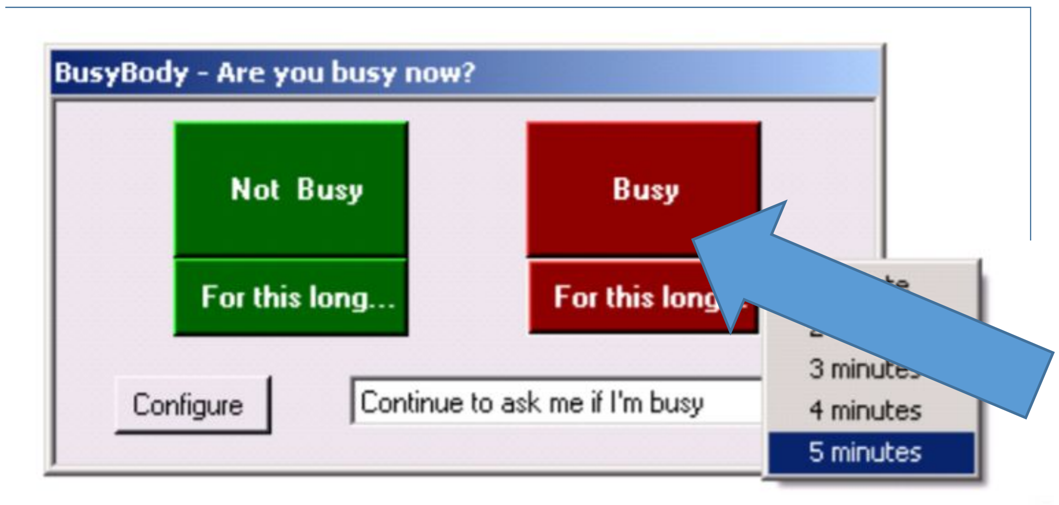


Every machine condition is different (machine states, machine installation, sensor location, operation conditions)

Example: Interruptibility

When there's a new notification to deliver, ask the following:





BusyBody, CSCW 2004

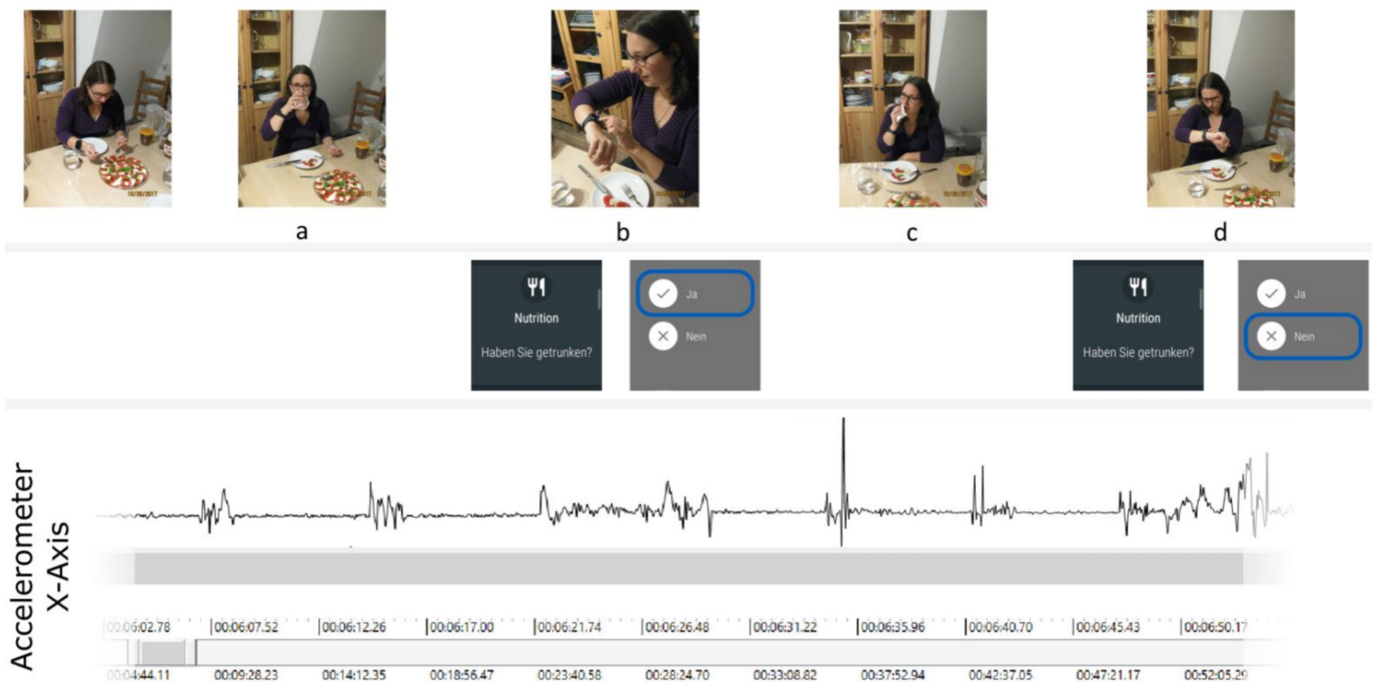
Example: Ringer Mode Change

Actively learning users' preferences for receiving audible notification preferences to automatically change sound volume

Using Decision-Theoretic Experience Sampling to Build Personalized Mobile Phone Interruption Models Stephanie Rosenthal, Anind K. Dey, Manuela Veloso, Pervasive 2011

Example: User Behavior Tracking

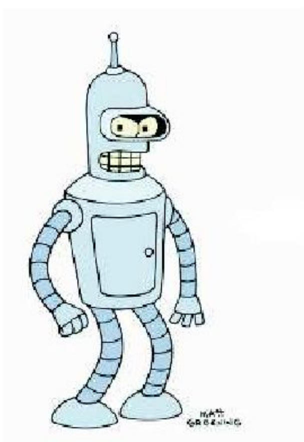
Find an interesting moment to ask questions! (Drinking Behaviors)



DrinkWatch: A Mobile Wellbeing Application Based on Interactive and Cooperative Machine Learning, DH 2018

(Passive) Supervised Learning

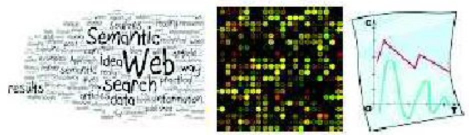
raw unlabeled data
$x_1, x_2, x_3, \ldots$



supervised learner
induces a classifier

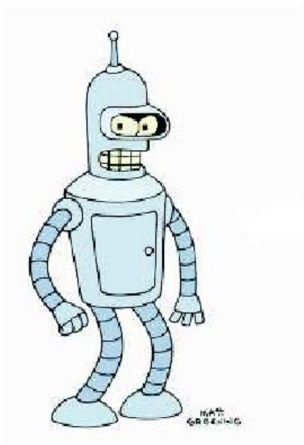
expert / oracle
analyzes experiments
to determine labels

(Passive) Supervised Learning



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

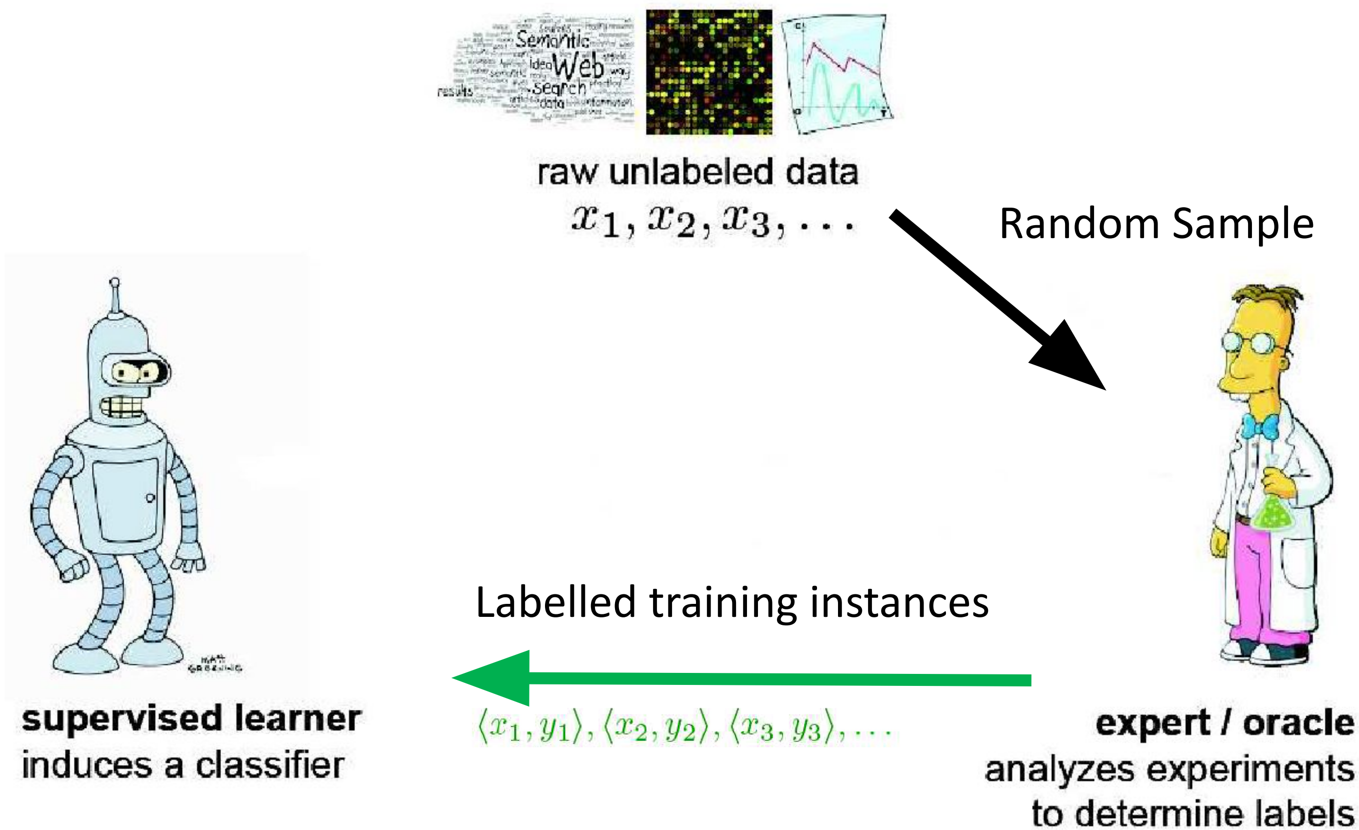
Random Sample



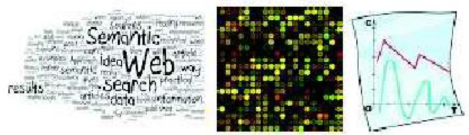
supervised learner
induces a classifier

expert / oracle
analyzes experiments
to determine labels

(Passive) Supervised Learning



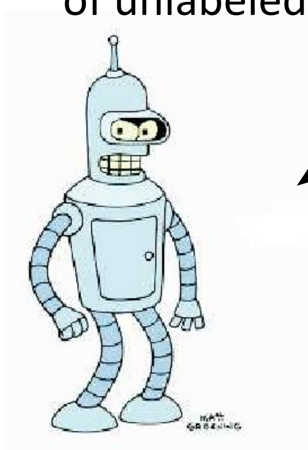
Semi-supervised Learning



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

Exploit the structure of unlabeled data

Random Sample



supervised learner
induces a classifier

Labelled training instances

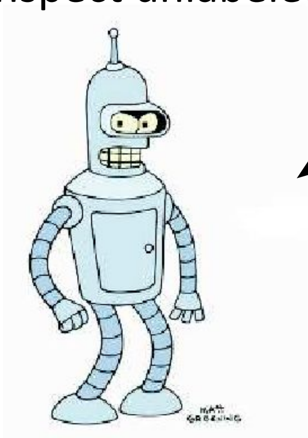
$\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \langle x_3, y_3\rangle, \ldots$

expert / oracle
analyzes experiments
to determine labels

Active Learning

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

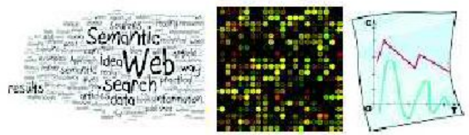
Inspect unlabeled data



supervised learner
induces a classifier

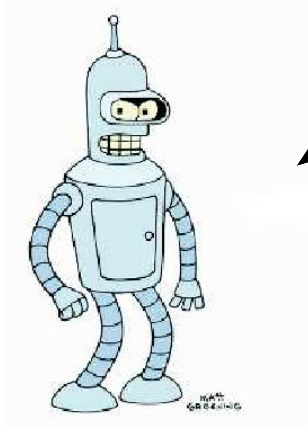
expert / oracle
analyzes experiments
to determine labels

Active Learning



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

Inspect unlabeled data

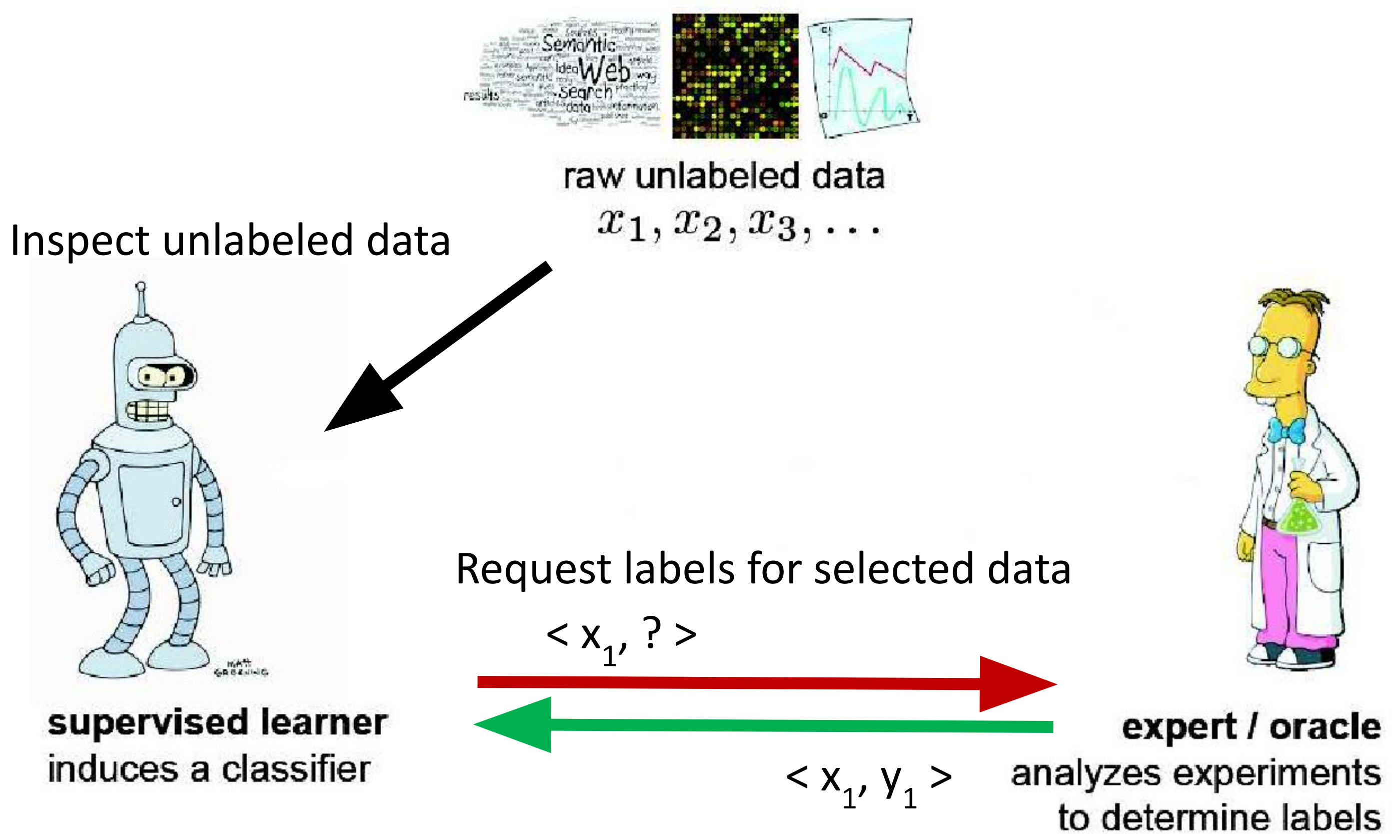


Request labels for selected data
$< x_1, ? >$

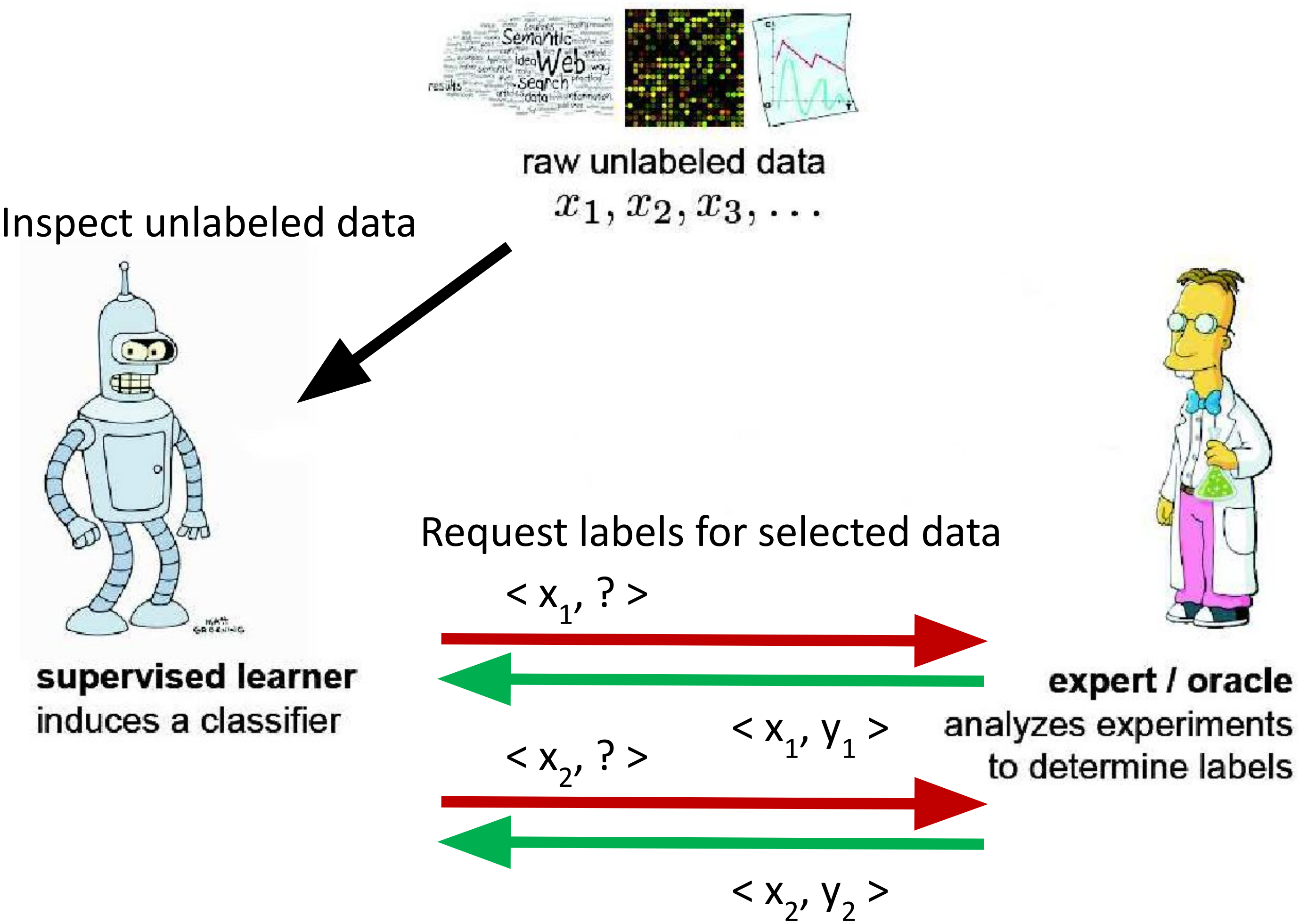
supervised learner
induces a classifier

expert / oracle
analyzes experiments
to determine labels

Active Learning



Active Learning



Active Learning vs Random Sampling

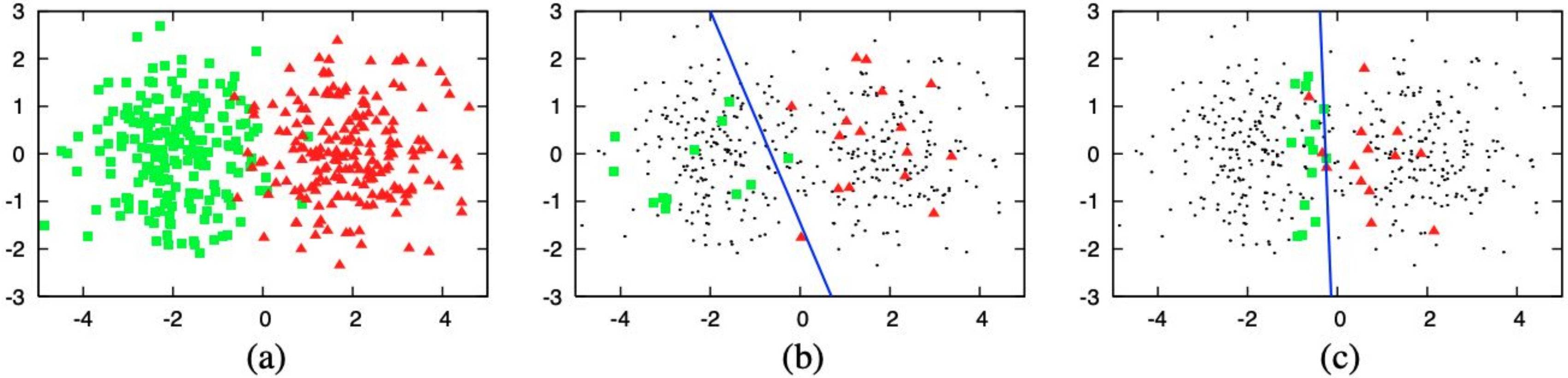
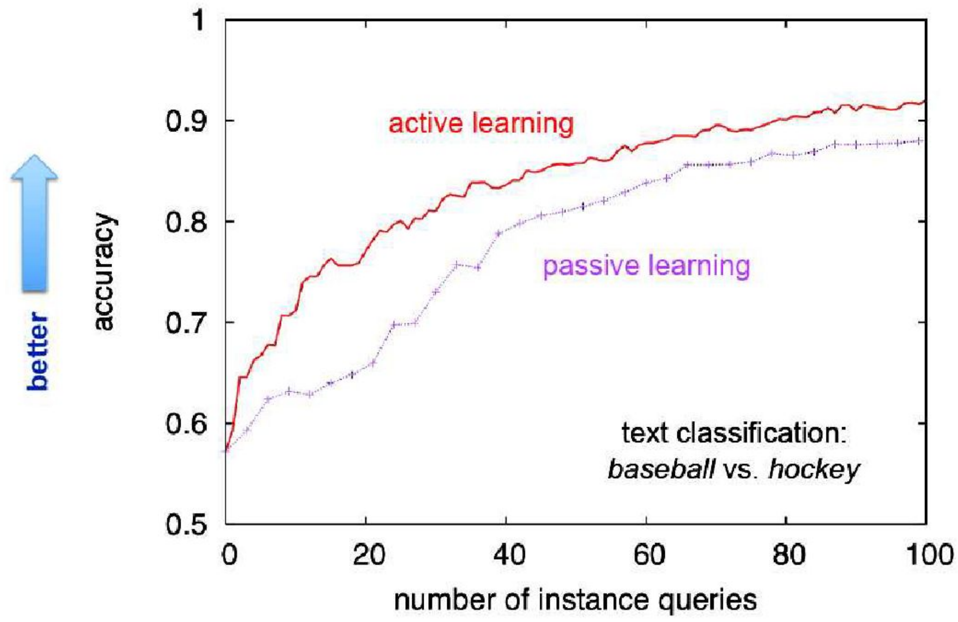


Figure 2: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem 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%).

Active Learning vs Random Sampling

- Passive Learning curve: Randomly selects examples to get labels for
- Active Learning curve: Active learning selects examples to get labels for

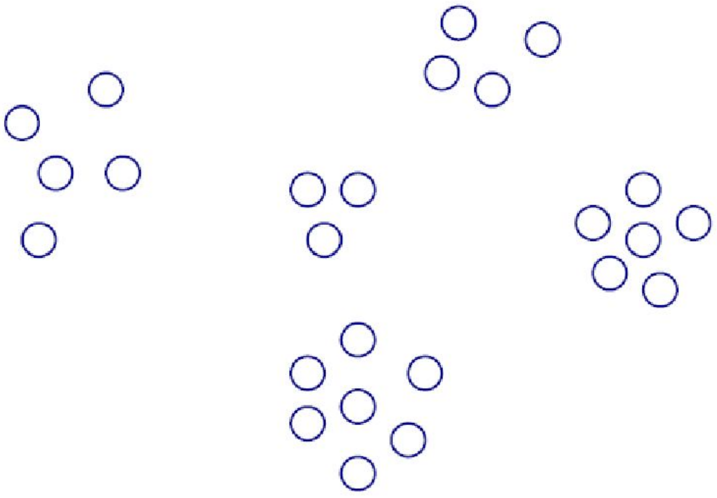
Learning Curves



A Naive Approach

- “random sampling” may work poorly in some cases

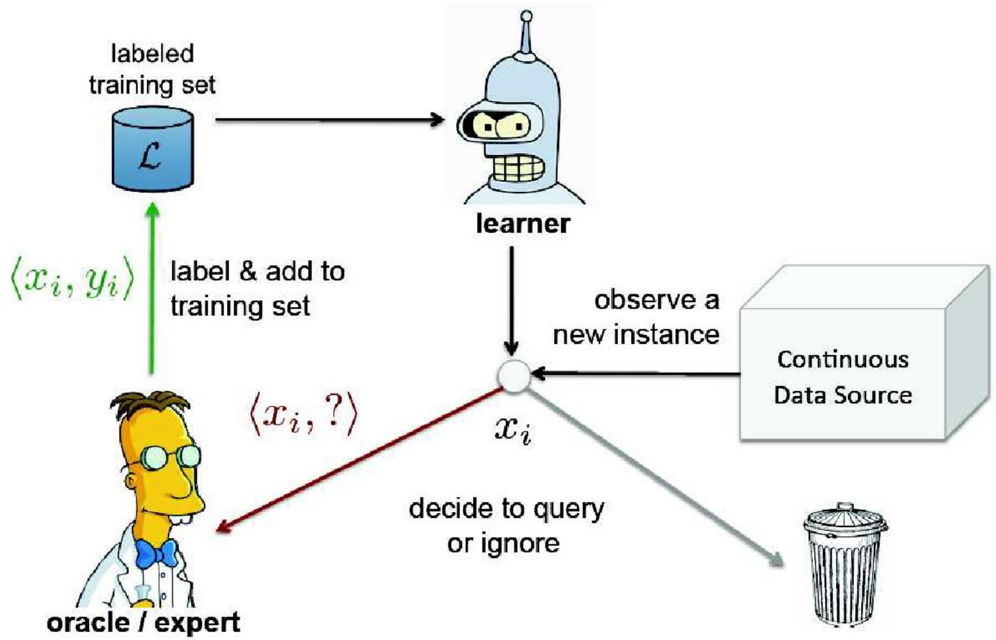
Suppose the unlabeled data looks like this.



Then perhaps we just need five labels!

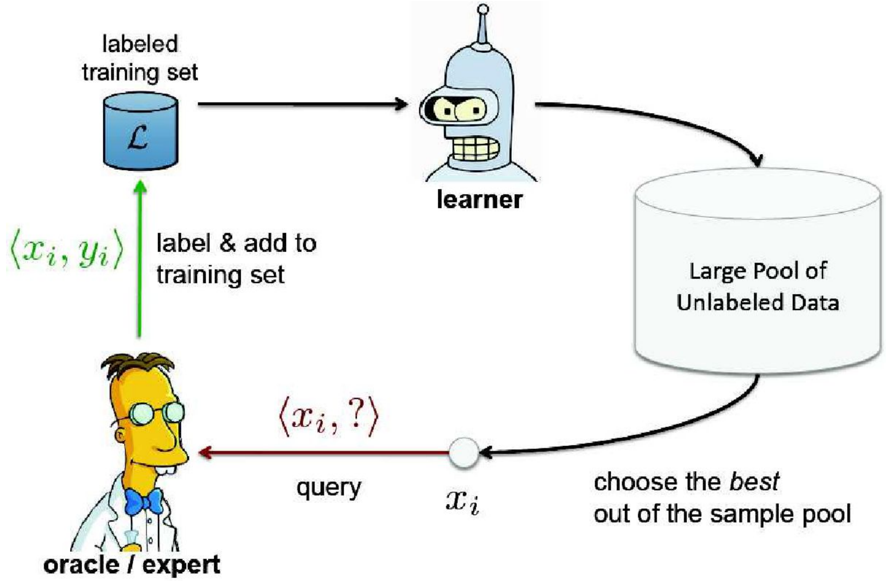
Types of Active Learning

- Largely falls into one of these two types:
  - **Stream-Based Active Learning**
    - Consider one unlabeled example at a time
    - Decide whether to query its label or ignore it



Types of Active Learning

- Largely falls into one of these two types:
  - **Pool-Based Active Learning**
    - Given: a large unlabeled pool of examples
    - Rank examples in order of informativeness
    - Query the labels for the most informative example(s)



Recap: How AL Operates?

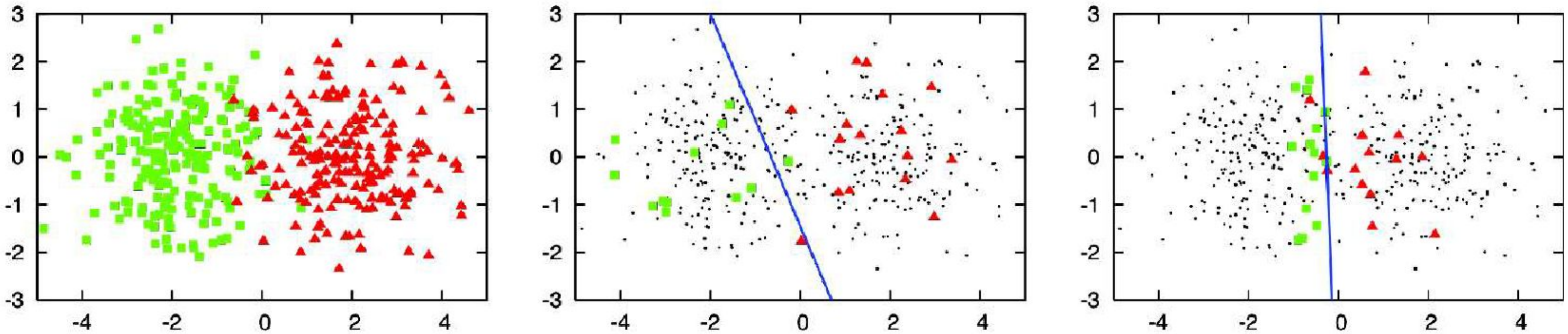
- Active Learning proceeds in rounds
- Each round has a current model (learned using the labeled data seen so far)
- Steps
  1. The current model is used to assess informativeness of unlabeled examples (e.g., using one of the query selection strategies)
  2. The most informative example(s) is/are selected
  3. The labels are obtained (by the labeling oracle)
  4. The (now) labeled example(s) is/are included in the training data
  5. The model is re-trained using the new training data
- The process repeat until we have budget left for getting labels

Query Selection Strategies

- Any Active Learning algorithm requires a query selection strategy
- Some examples:
  - Uncertainty Sampling
  - Query By Committee (QBC)
  - Density Weighted Methods
  - Expected Model Change
  - Expected Error Reduction
  - Variance Reduction

Uncertainty Sampling

- Select examples which the current model θ is the **most uncertain** about
- Various ways to **measure uncertainty**. For example:
  - Based on the distance from the hyperplane
  - Using the label probability $P_\theta(y|x)$ (for probabilistic models)



400 instances sampled from 2 class Gaussians

random sampling
30 labeled instances
(accuracy=0.7)

uncertainty sampling
30 labeled instances
(accuracy=0.9)

Uncertainty sampling based on the distance from the hyperplane (i.e., margin based)

Uncertainty Sampling

- Select examples which the current model θ is the **most uncertain** about
- Various ways to **measure uncertainty**. For example:
  - Based on the distance from the hyperplane
  - Using the label probability $P_\theta(y|x)$ (for probabilistic models)
- Some typically used measures based on label probabilities:

Least Confident: $x^*_{LC} = \text{argmax}_x\, 1 - P_\theta(\hat{y}|x)$
where $\hat{y}$ is the most probable label for x under the current model θ

Smallest Margin: $x^*_{SM} = \text{argmin}_x\, P_\theta(y_1|x) - P_\theta(y_2|x)$
y_1, y_2 are the two most probable labels for x under the current model

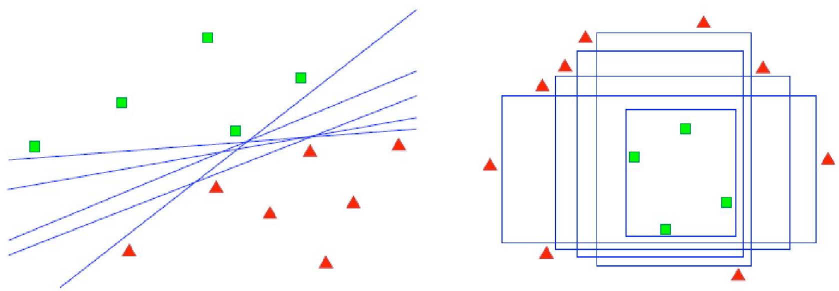
Label Entropy: choose example whose label entropy is maximum

$$x^*_{LE} = \underset{x}{\text{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

where y_i ranges over all possible labels

Query-By-Committee (QBC)

- QBC uses a committee of models $C = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$
- All models trained using the currently available labeled data L



Version space examples for (a) linear and (b) axis-parallel box classifiers

All hypotheses are consistent with the labeled training data in L (as indicated by shaded polygons), but each represents a different model in the version space

- **How is the committee constructed?** Some possible ways:
  - Sampling different models from the model distribution $P(\theta|L)$
  - Using ensemble methods (query-by-bagging/boosting, etc.)
- All models vote their predictions on the unlabeled pool
- **The example(s) with maximum disagreement is/are chosen for labeling**
- Each model in the committee is re-trained after including the new example(s)

Query-By-Committee (QBC)

- Measuring disagreement is the Vote Entropy
  - Vote Entropy

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

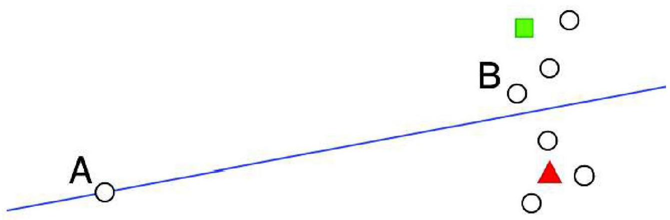
    - y_i ranges over all possible labels
    - $V(y_i)$: number of votes that a label y_i received from the committee members' predictions
    - C: committee size
  - Kullback-Leibler (KL) divergence

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta^{(c)}} \| P_{\mathcal{C}}), \quad D(P_{\theta^{(c)}} \| P_{\mathcal{C}}) = \sum_i P_{\theta^{(c)}}(y_i|x) \log \frac{P_{\theta^{(c)}}(y_i|x)}{P_{\mathcal{C}}(y_i|x)}$$

$\theta^{(c)}$ represents a particular model in the committee, and C represents the committee as a whole, thus $P_C(y_i \mid x)$ is the "consensus" probability that y_i is the correct label

Effect of Outlier Examples

- Uncertainty Sampling or QBC may wrongly think an outlier to be an informative example
- Such examples won't really help (and can even be misleading)



- Other robust query selection methods exist to deal with outliers
- Instead of using the confidence of a model on an example, see how a labeled example affects the model itself (various ways to quantify this)
- The example(s) that affects the model the most is probably the most informative

Other Query Selection Methods

- Density Weighting
  - Informative instances should not only be those which are **uncertain**, but also those which are "**representative**" of the underlying distribution (i.e., staying in the dense regions of the input space)
  - Weight the informativeness of an example by its average similarity to the entire unlabeled pool of examples
- Expected Model Change
  - Select the example whose inclusion brings about **the maximum change in the model** (e.g., the gradient of the loss function w.r.t. the parameters)
  - Here, we don't know the true label for each query instance, so we approximate using expectation over all possible labels under the current model θ
- Expected Error Reduction
  - Select example that reduces the **expected generalization (future) error** the most, which is measured w.r.t. the **remaining unlabeled examples** (using the expected labels)
- Variance Reduction
  - Select example(s) that reduces the **model variance** by the most

Practical Issues – Batching & Cost

- Batch-Mode Active Learning
  - Batch-mode active learning allows the learner to query instances in groups, which is better suited to parallel labeling environments or models with slow training procedures
- Variable Labeling Costs
  - There is variance not only in label quality from one instance to the next, but also in the cost of obtaining that label
  - Cost-sensitive active learning approaches explicitly accounts for varying label costs while selecting queries

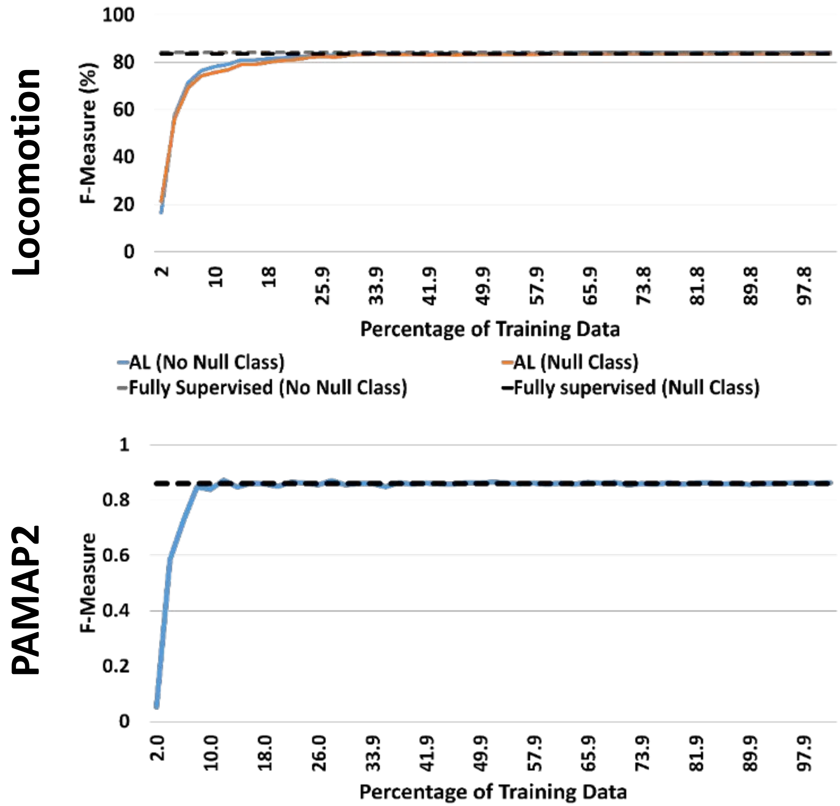
Practical Issues – Tasks & Stopping

- Multi-Task Active Learning
  - A single query will be labeled for multiple tasks, and attempt to assess the informativeness of a query with respect to all the learners involved
- Stopping Criteria:
  - Cost of acquiring new training data is greater than the cost of the errors made by the current model
  - Recognize when the accuracy of a learner has reached a plateau, and acquiring more data is likely a waste of resources
  - Examples:
    - Active learning ceases to be useful once that measure begins to level-off or degrade
    - Measures of stability, self-confidence within the learner, or information gain of adding new samples (IMWUT’19)

Practical Issues – Stopping

- OPPORTUNITY Dataset
  - 10-20% of data for locomotion recognition (both including and excluding Null Class)
- PAMAP2 Dataset
  - 8% of data to reach fully supervised performance

Offline AL: Pool-based Approach



Leveraging Active Learning and Conditional Mutual Information to Minimize Data Annotation in Human Activity Recognition, IMWUT 2019

Practical Issues – Noisy Oracle

- Noisy Oracle:
  - Another strong assumption in most active learning work is that the quality of labeled data is high
  - If labels come from an empirical experiment (e.g., in biological, chemical, or clinical studies), then one can usually expect some noise to result from the instrumentation of experimental setting
  - Even if labels come from human experts, they may not always be reliable, for several reasons
    - Some instances are implicitly difficult for people and machines
    - People can become distracted or fatigued over time, introducing variability in the quality of their annotations

Practical Issues – Noisy Oracle

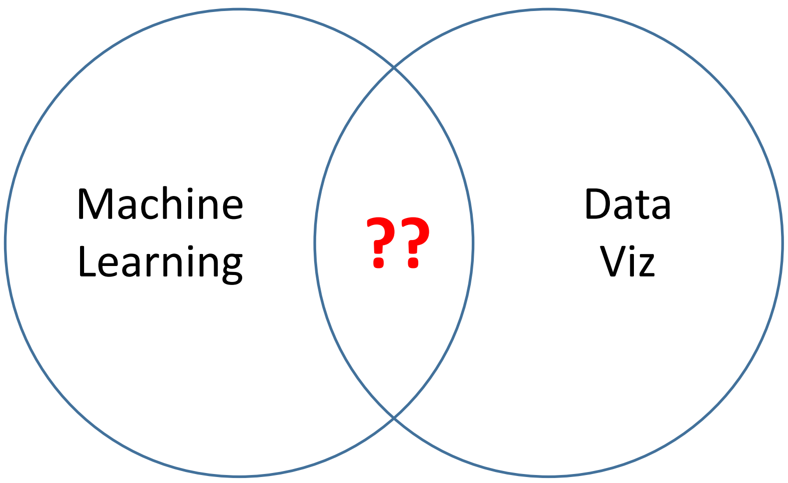
- Noisy Oracle:
  - Humans are not “oracles” – prone to make errors or produce poor quality work
  - How can we build systems that are robust against such errors or low quality work?
    - Avoid errors – help users to label correctly, ensure correctness (interactive visualization, and guidance)
    - Eliminate errors afterward – let the users (or others) to re-review and filter errors; or collect multiple samples to “average out” errors (e.g., via M-Turk)
    - Use robust models – use ML models robust against potential outliers

Practical Issues – Ask timing

- Ask Timing
  - Learner should decide when to query as opposed to letting the user choose an example
    - Predefine Conditions? Contexts? Active learning?
  - Some tasks require "experience sampling" (due to recall bias – difficult to answer later)
  - Batch answering could be okay depending on "recallability"
    - Minutes: emotion, attention
    - Hours: mood, rough activity
    - Day: rough activity
  - Contextual cues lead to better judgement
    - Examples: objective videos, or contextual summary

Practical Issues – VIL

- Data-driven knowledge discovery
  - Combine the strength of Machine Learning and Visualization



VIAL: a unified process for visual interactive labeling, Bernard et al., VIS 2017

Practical Issues – VIL

- **Active learning : Model-based strategy**
  - Algorithms select instances to improve a learning model best
  - Querying labels for these instances from an oracle (or user)
- **Visual-interactive labeling : User-based strategy**
  - Visual-interactive interfaces enable users to select instances
  - Exploration as a means to identify meaningful instances

Towards rapid interactive machine learning: evaluating tradeoffs of classification without representation, IUI 2019

Summary

- Active Learning: Label efficient learning strategy
- Based on judging the informativeness of examples
- Several variants possible. E.g.,
  - Different examples having different labeling costs
  - Access to multiple labeling oracles (possibly noisy)
  - Active Learning on features instead of labels (e.g., if features are expensive)
- Being “actively” used in industry (IBM, Microsoft, Google)