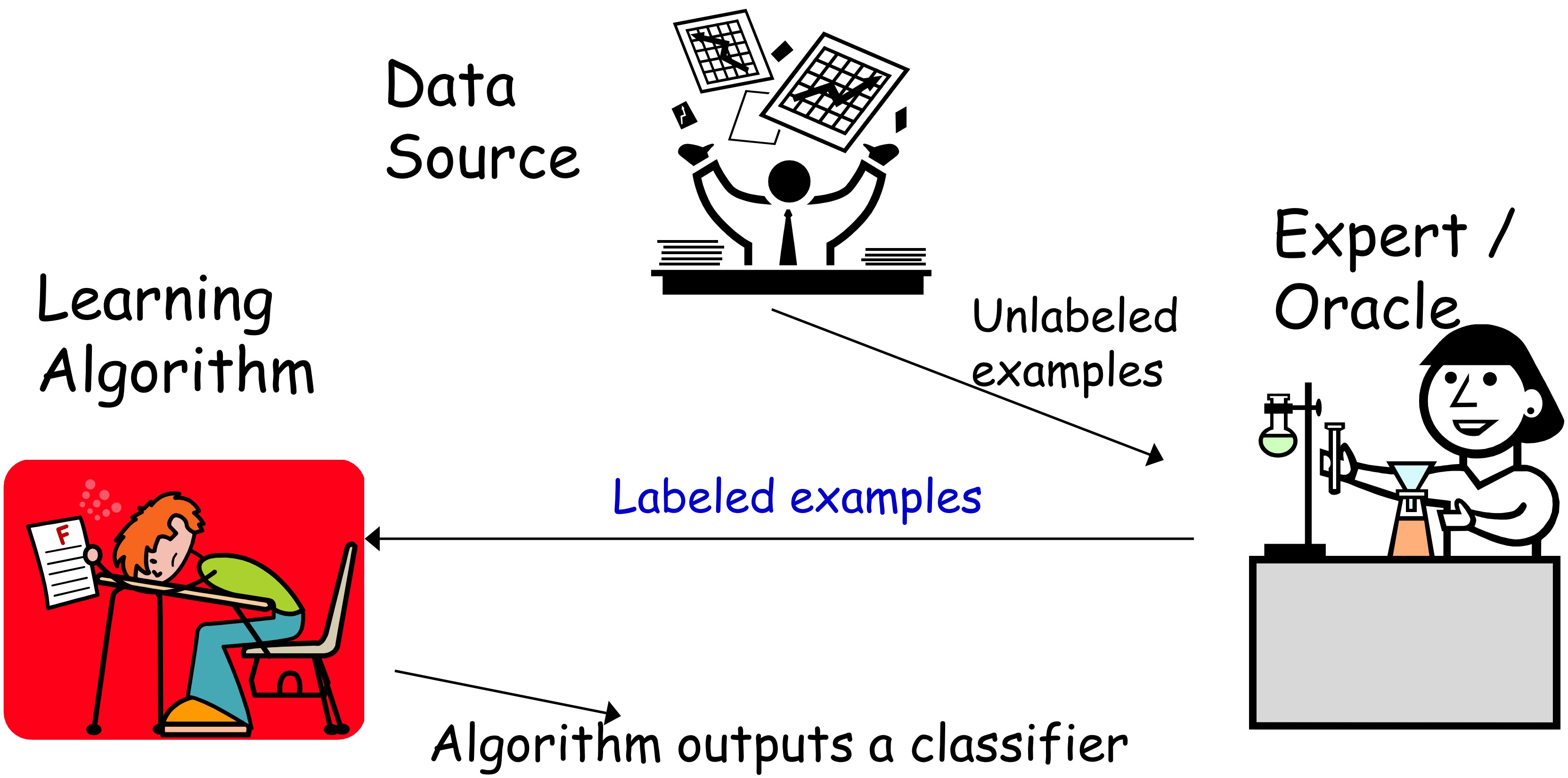
Robust Interactive Learning

Maria-Florina Balcan
Georgia Institute of Technology

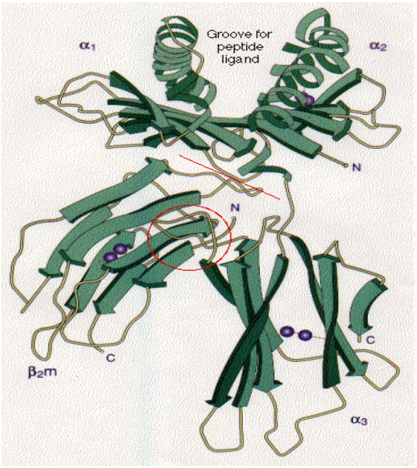
Joint with Steve Hanneke

Classic Statistical Learning Paradigm: passive supervised learning (labeled data only)

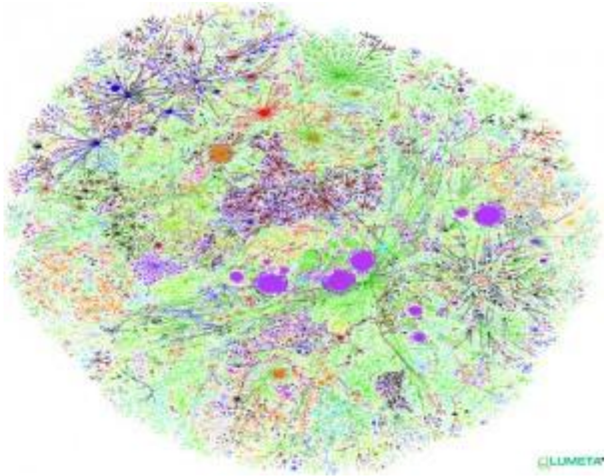


Massive Amounts of Raw Data

Only a tiny fraction can be annotated by human experts.



Protein sequences

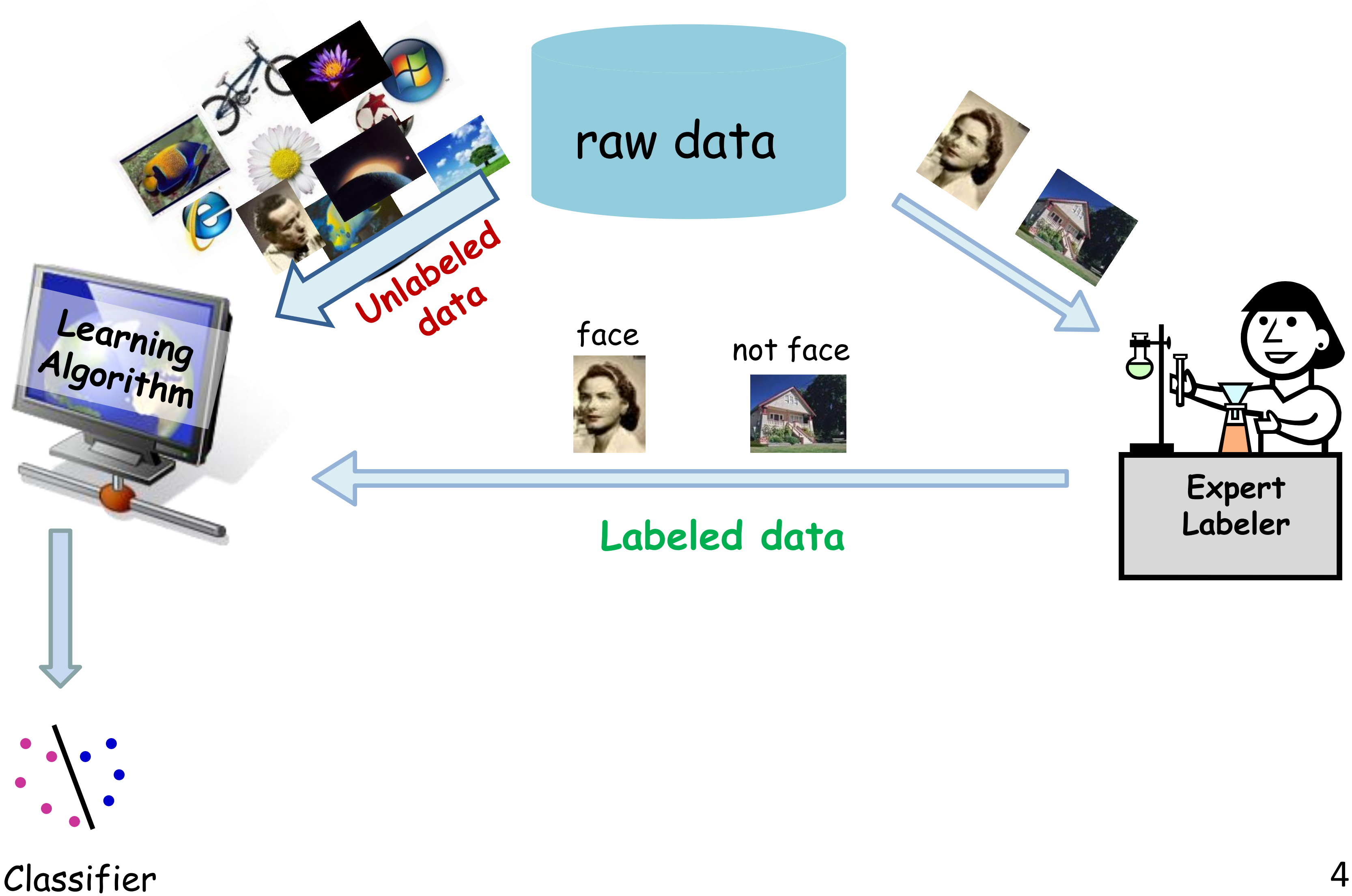


Billions of webpages

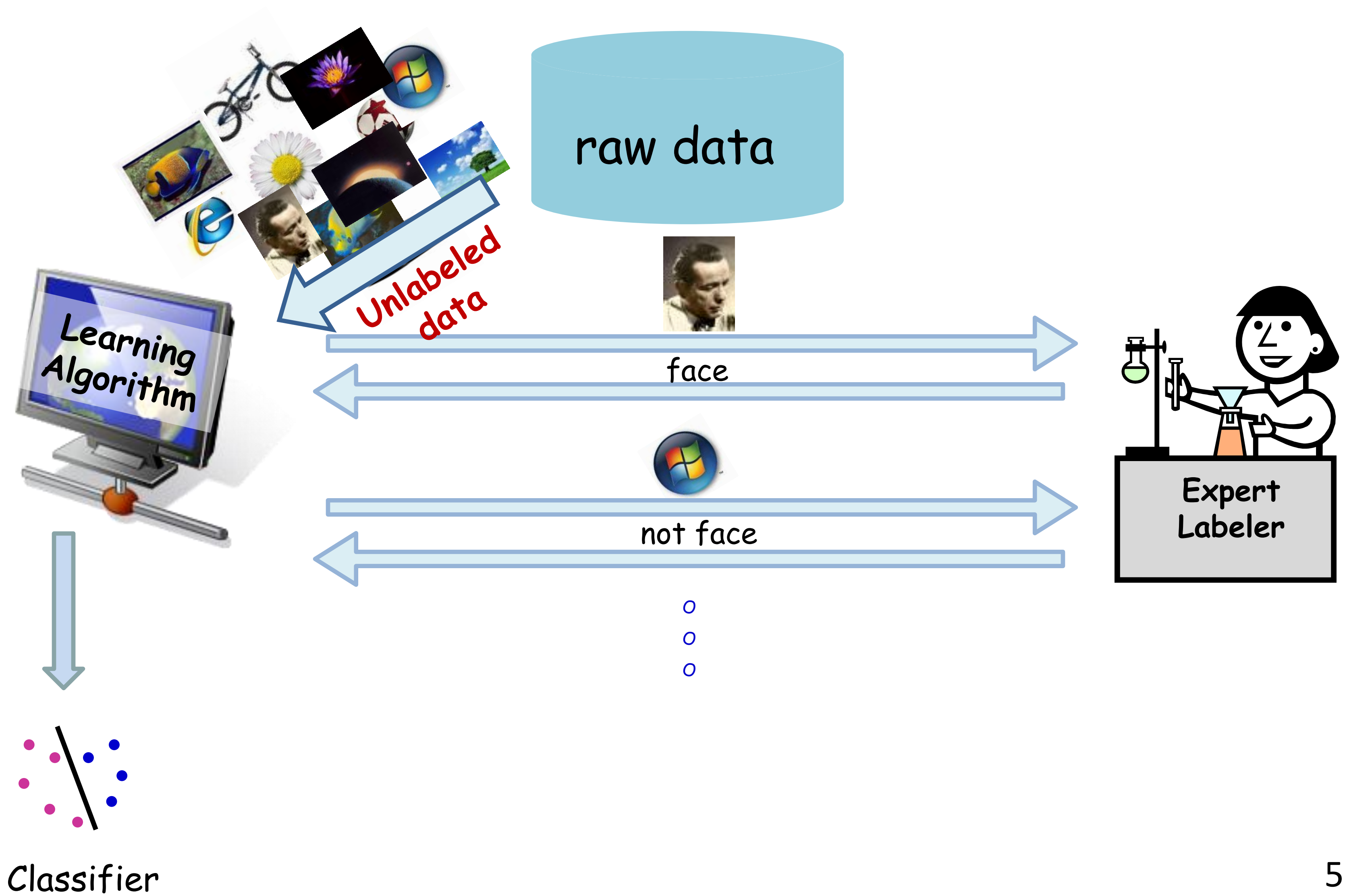
Images

Major Efforts Recently: Unlabeled Data and Interaction for Learning

Semi-Supervised Learning



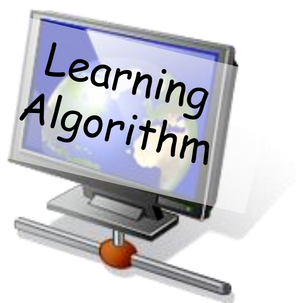
Active Learning



Important direction: richer interactions with the expert.

Better Accuracy

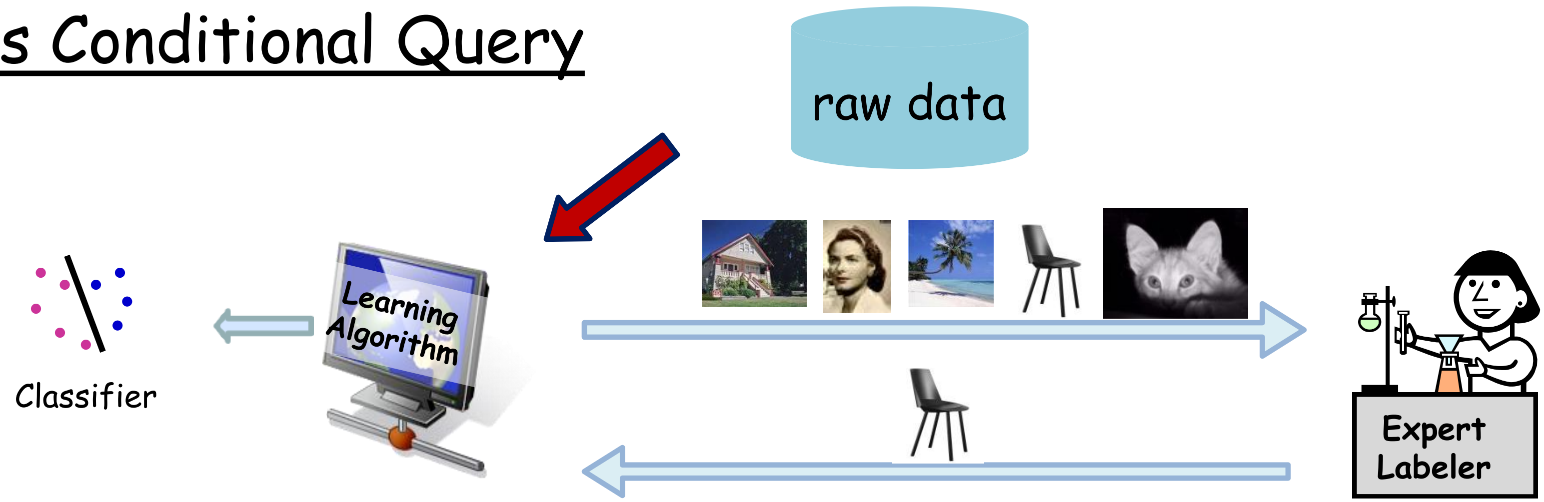
Fewer queries



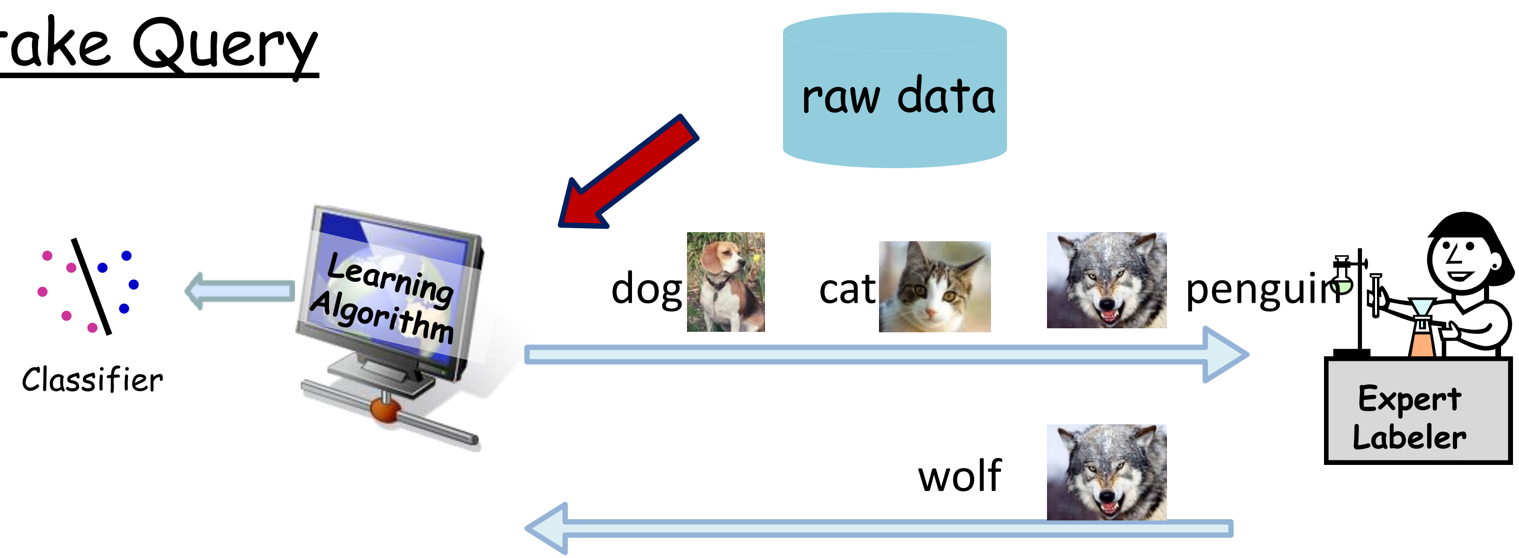
Natural interaction

New Types of Interaction

Class Conditional Query

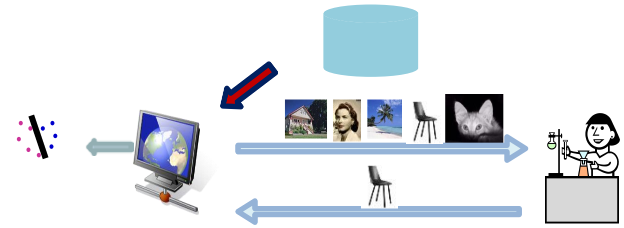


Mistake Query



Class Conditional & Mistake Queries

- Used in practice, e.g. Faces in IPhoto.
- Lack of theoretical understanding.



- Realizable (Folklore): much fewer queries than label requests.

This work:

Tight bounds on the # of CCQs to learn with noise (agnostic, bounded noise, one-sided noise).

Preliminary results for more general queries.

Formal Model

- X instance space, $Y = \{1,2, \dots, k\}$ label space.
- $D_{\{X,Y\}}$ fixed target distribution, $D_{\{X\}}$ marginal over X.
- An i.i.d sequence (x_1, y_1), $(x_2, y_2), (x_3, y_3), \dots$, each with distr $D_{\{X,Y\}}$.

- Algo can access x_i values; info about y_i obtained via CCQs.
  CCQ($S = \{x_{i_1}, \dots, x_{i_m}\}$, label l)
  - If $y_{i_j} \neq l$, for all j, expert says "none";
  - otherwise, returns an arbitrary x_{i_j} s.t. $y_{i_j} = l$

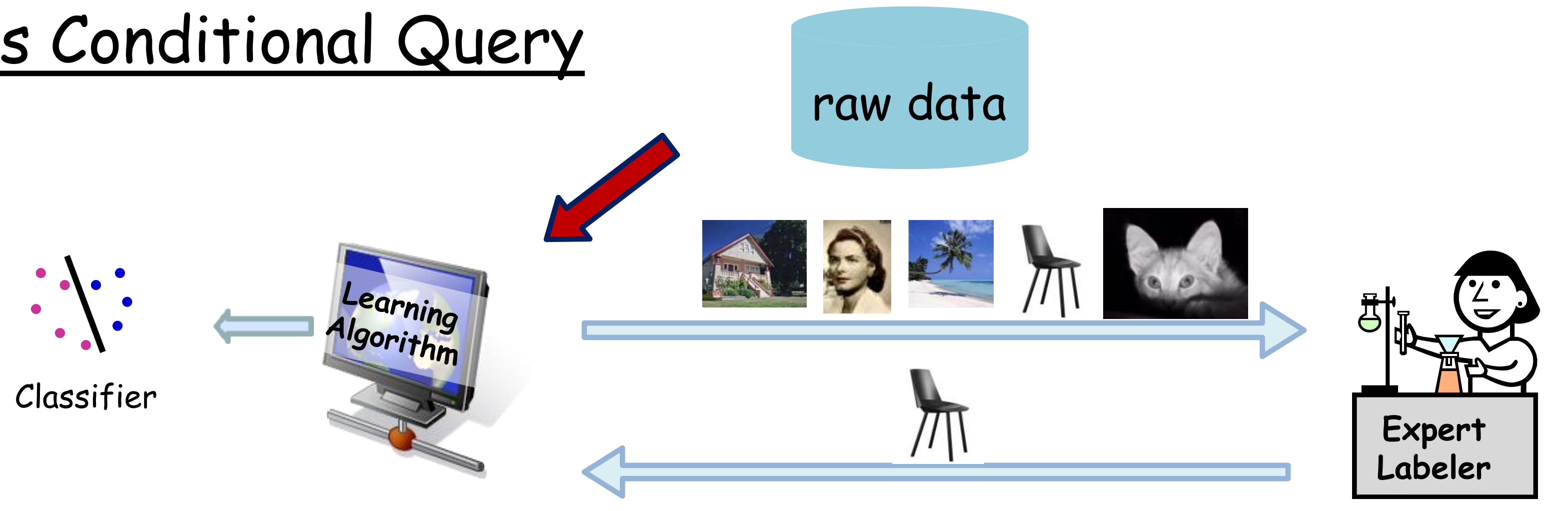
- FindMistake(S,h) based on k CCQs
  - For each l, query the set $\{x \in S: h(x) \neq l\}$ for label l
  - If received back an example (x, l), return (x, l)

- C, Ndim$(C) = d$ (Natarajan dimension)

  largest m s.t. $\exists\, (a_1, b_1, c_1), \dots, (a_m, b_m, c_m) \in X \times Y \times Y$, $b_i \neq c_i$ s.t.
  $\{b_1, c_1\} \times \cdots \times \{b_m, c_m\} \subseteq \{h(a_1), \dots, h(a_m)\}$

Class Conditional Query

Class Conditional Query



Folklore Result (Realizable Case):

$\mathrm{kd}\log\left(\frac{\mathrm{k}}{\epsilon}\right)$ queries sufficient to learn an ϵ – accurate classifier, assuming that the target belongs to C of dimension d.

- Run halving algorithm

Agnostic Case

$\mathrm{Agnostic}(C, \eta) = \{D_{X,Y}: \inf_{h \in C} \mathrm{err}(h) \leq \eta\}$

η Passive QC

For any class C, $\mathrm{QC}\left(\epsilon, \delta, C, \mathrm{Agnostic}(C, \eta)\right) = \tilde{\Theta}\left(d\left(\frac{\eta}{\epsilon}\right)^2\right)$ [Const k]

Upper bound: $\tilde{O}\left(kd\left(\frac{\eta}{\epsilon}\right)^2\right)$

Phase 1: Robust halving algorithm

Output h, $\mathrm{err}(h) \leq 10(\eta + \epsilon)$; only $\tilde{O}\left(kd\log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Phase 2: A simple refining algorithm

$\tilde{O}\left(kd\left(\frac{\eta}{\epsilon}\right)^2\right)$ queries to turn h into one of error $\eta + \epsilon$

Agnostic Case, Upper Bound

Phase 1: Output h of error $10(\eta+\epsilon)$ with $\tilde{O}\left(kd\log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Input $U=\left(x_1,x_2,\ldots,x_{ps}\right)$, V- ϵ-cover. $s=\frac{1}{16\eta}$, $N=\log\left(\frac{\log|V|}{\delta}\right)$; $b=1$

While b.

- Draw $S_1,\ldots,S_N$ of size s uniformly from U.
- For each i, Find-Mistake (S_i, $\mathrm{plur}(V)$). If mistake record it $(\tilde{x}_i,\tilde{y}_i)$.
- If mistakes on more than $N/3$ sets, remove from V every h that makes a mistake on more than $N/9$ examples $(\tilde{x}_i,\tilde{y}_i)$; else $b=0$.

Output $\mathrm{plur}(V)$.

I.e., eliminate h when makes at least one mistake in some number out of several sets chosen at random from U.

- rather than eliminating h for a mistake, as in halving.

Agnostic Case, Upper Bound

Phase 1: Output h of error $10(\eta + \epsilon)$ with $\tilde{O}\left(kd \log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Input $U = (x_1, x_2, \ldots, x_{ps})$, V- ϵ-cover. $s = \frac{1}{16\eta}$, $N = \log\left(\frac{\log|V|}{\delta}\right)$; $b = 1$

While b.

- Draw $S_1, \ldots, S_N$ of size s uniformly from U.
- For each i, Find-Mistake $(S_i, \mathrm{plur}(V))$. If mistake record it $(\tilde{x}_i, \tilde{y}_i)$.
- If mistakes on more than $N/3$ sets, remove from V every h that makes a mistake on more than $N/9$ examples $(\tilde{x}_i, \tilde{y}_i)$; else $b = 0$.

Output $\mathrm{plur}(V)$.

Proof Sketch

- Best h in V survives (does not make mistakes on too many sets).
- If $\mathrm{err}(\mathrm{plur}(V)) \geq 10\,\eta$, then $\mathrm{plur}(V)$ make mistakes on enough sets; so, a constant fraction of V will make mistakes on more sets than the best classifier & eliminate a constant fraction of V.

Upper Bound, Agnostic Case

$\mathrm{Agnostic}(C, \eta) = \{D_{X,Y}: \inf_{h \in C} \mathrm{err}(h) \leq \eta\}$

η Passive QC

For any class C, $\mathrm{QC}(\epsilon, \delta, C, \mathrm{Agnostic}(C, \eta)) = \tilde{\Theta}\left(d\left(\frac{\eta}{\epsilon}\right)^2\right)$ [Const k]

Phase 1: Robust halving algorithm

Output h, $\mathrm{err}(h) \leq 10(\eta + \epsilon)$; only $\tilde{O}\left(kd \log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Phase 2: A simple refining algorithm

$\tilde{O}\left(kd\left(\frac{\eta}{\epsilon}\right)^2\right)$ queries to turn h into one of error $\eta + \epsilon$

Pick $\tilde{O}\left(d \frac{\eta}{\epsilon^2}\right)$ unlabeled samples, and repeatedly run FindMistake to find all the mistakes h makes.

We recover the true labels of this unlabeled sample and run ERM.

Agnostic Case, Lower Bound

For $0 < 2\epsilon \le \eta \le \frac{1}{4}$, $\mathrm{QC}\left(\epsilon, \frac{1}{4}, \mathrm{C}, \mathrm{Agnostic}(\mathrm{C}, \eta)\right) = \Omega\left(\mathrm{d}\left(\frac{\eta}{\epsilon}\right)^2\right)$

Proof Sketch: Reduction from (binary)AL to multiclass CCQs for AL hard case.

- AL hard case [BDL'09].

$(\mathrm{x}_0, \mathrm{y}_0, \mathrm{z}_0), \ldots, (\mathrm{x}_{d-1}, \mathrm{y}_{d-1}, \mathrm{z}_{d-1})$

$\mathbf{x_0}$	x_1	…	x_{d-1}
$1-\beta$	β/d-1		β/d-1

$$\gamma = \frac{2\epsilon}{\beta} = \frac{\epsilon}{\eta + 2\epsilon}$$

  - At x_0, $\mathrm{Y} = \mathrm{y}_0$. At x_i, is $\mathrm{Y} = \mathrm{z}_\mathrm{i}$ with prob. $1/2 + \gamma \mathrm{b}_\mathrm{i}$ and $\mathrm{Y} = \mathrm{y}_\mathrm{i}$ with prob. $1/2 - \gamma \mathrm{b}_\mathrm{i}$.
  - BDL'09 : set b_i s. t. any algo makes $\Omega\left(\mathrm{d}\left(\frac{\eta}{\epsilon}\right)^2\right)$ AL queries for $\delta = 1/2$.
- Convert a CCQ algo into an AL algo for this case.

CCQ(S, l)

S

R

(x_{i_1}, y_{i_1}) (x_{i_2}, y_{i_2}) … (x_{i_s}, y_{i_s})

  - answer CCQs using $\le 1/(1/2 - \gamma)$ AL requests (for label y_i, $\mathrm{i} \ge 1$, $\frac{1}{2} - \gamma$ prob. to get desired label).

Bounded Noise

Bounded Noise: $\mathrm{BN}(\mathrm{C}, \alpha) = \{D_{X,Y} \colon \exists h^* \in C, P_{D_{X,Y}}[Y \neq h^*(X) | X] \leq \alpha\}$

For any class C, $\mathrm{QC}(\epsilon, \delta, \mathrm{C}, \mathrm{BN}(\mathrm{C}, \alpha)) \approx \alpha\, \mathrm{AL\_QC}(\epsilon, \delta, \mathrm{C}, \mathrm{BN}(\mathrm{C}, \alpha))$

Lower bound:

- reduction from multi-class AL (as for the agnostic case)

Upper bound: many existing AL algos batch based

- reduction to multi-class (batch based) AL algos

Bounded Noise, Upper Bound

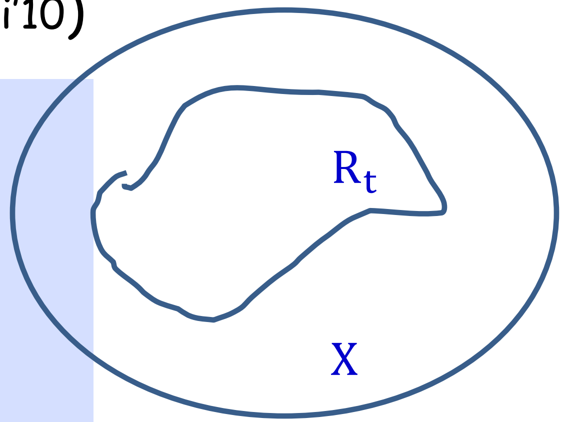
For any class C, $\mathrm{QC}(\epsilon, \delta, \mathrm{C}, \mathrm{BN}(\alpha)) \approx \alpha\, \mathrm{AL\_QC}(\epsilon, \delta, \mathrm{C}, \mathrm{BN}(\alpha))$

Many existing AL algos batch based

A^2 (Balcan, Beygelzimer, Langford'06), K's algo (Koltchinski'10)

For t= 1,2,…

Region $\mathrm{R_t}$, m &expects back m labeled examples from the cond. distr. given R.



E.g, in A^2, $\mathrm{V_t}$ set of surviving classifiers, $\mathrm{R_t}$ = DIS($\mathrm{V_t}$)

Bounded Noise, Upper Bound

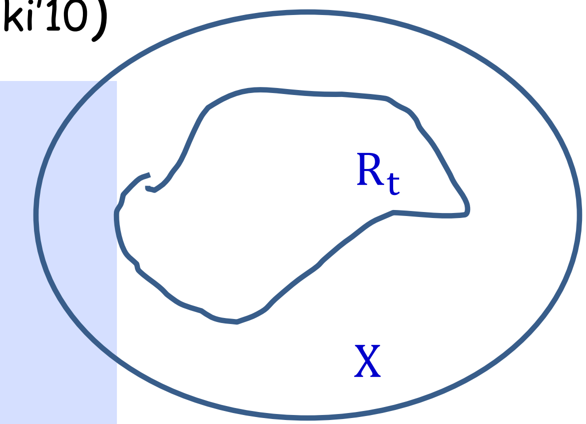
For any class C, $\mathrm{QC}(\epsilon, \delta, \mathrm{C}, \mathrm{BN}(\alpha)) \approx \alpha\, \mathrm{AL\_QC}(\epsilon, \delta, \mathrm{C}, \mathrm{BN}(\alpha))$

Many existing AL algos batch based

A^2 (Balcan, Beygelzimer, Langford'06), K's algo (Koltchinski'10)

For t= 1,2,…

Region $\mathrm{R_t}$, m &expects back m labeled examples from the cond. distr. given R.



Use our agnostic algo to only make $\alpha \cdot m$ CCQ's

Conclusions, Open Questions

Nearly tight upper and lower bounds on the #of CCQs for the agnostic, bounded noise, one sided noise.

Open Questions

- Efficient algorithms for CCQs.
- Tight bounds for general queries.

Other General Abstract Queries

- Queries that allow to prune down the set of consistent labellings [Balcazar et al 2010].

  E.g., ask for two examples of opposite labels within a distance of each other in a set S, and for their labels.

- Bounds in terms of IAdim, related to the complexity of verification (generalization of teaching dimension for label requests)

  - IAdim(f,V,U) the smallest #of queries s. t. any valid answers consistent with f will leave at most one equivalence class in V[U] consistent with answers.
  - when target in V, number of queries to verify it is the target (in a non-adaptive way).

Other General Abstract Queries

For any C, $\mathrm{QC}\big(\epsilon, \delta, \mathrm{C}, \mathrm{Agnostic}(\mathrm{C}, \eta)\big) = \mathrm{O}\left(\mathrm{IAdim}\left(\frac{\eta}{\epsilon}\right)^2 \mathrm{d}\log\frac{1}{\epsilon}\right)$

<u>Proof Sketch</u>

- Replace FindMistake(S,h,V) with the queries in the def. of IAdim.
- If there exists h in V that coincides with the true labels this behaves like FindMistake.
- Take small enough sample so that there is a function in the class with this property.