SafePredict:
A Machine Learning Meta-Algorithm That Uses Refusals to Guarantee Correctness

David Ramirez (dard@princeton.edu),
Mustafa A. Kocak,
Elza Erkip,
Dennis E. Shasha

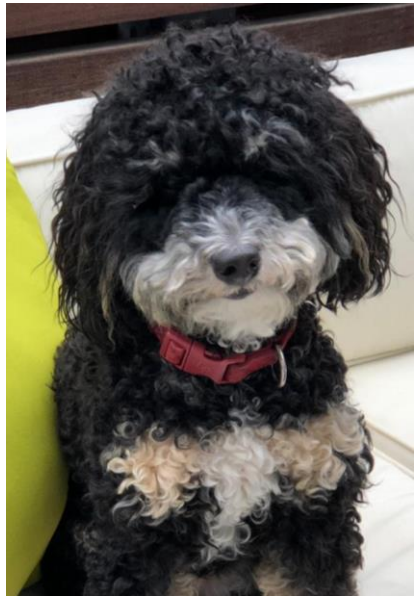
PRINCETON UNIVERSITY

BROAD INSTITUTE

NEW YORK UNIVERSITY

Correctness is necessary for algorithms

Observation



Expert Prediction

Observation

Expert Prediction

Reliability is crucial in risk-critical applications!

e.g.,   

Increasing Correctness

Observation

Crowd of experts can outperform one expert!

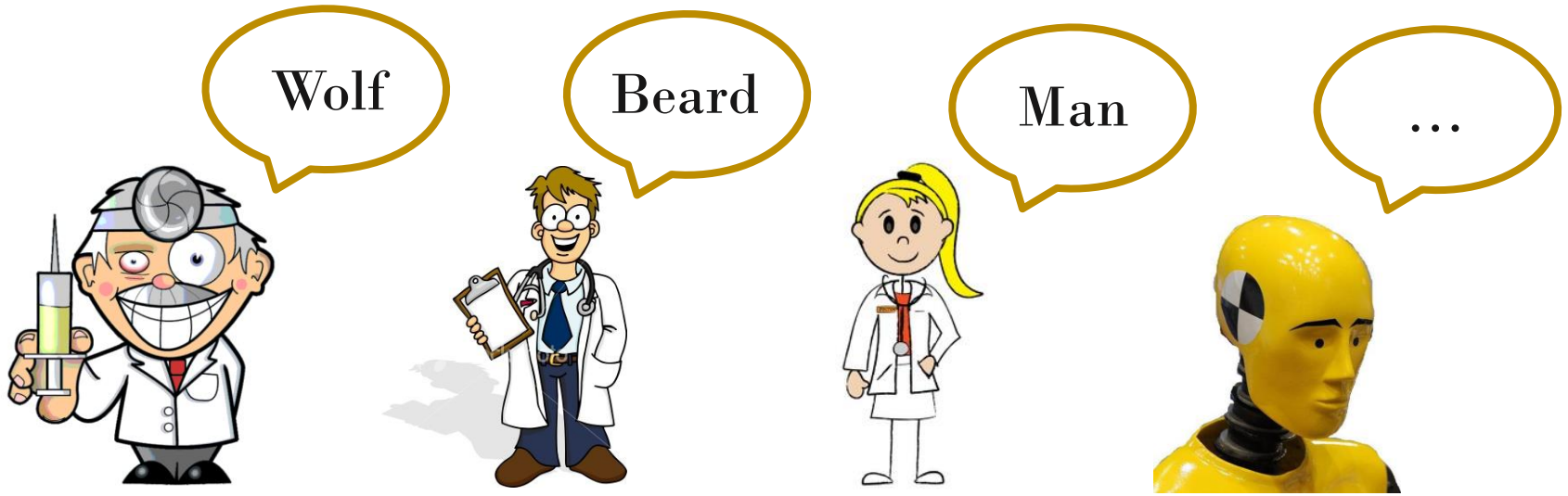
Goal: Predict the label of an observed object within a given target error rate

Problem: Unless a lot of assumptions are made, hard to meet error rate!

Correctness Guarantees via Refusals

Observation

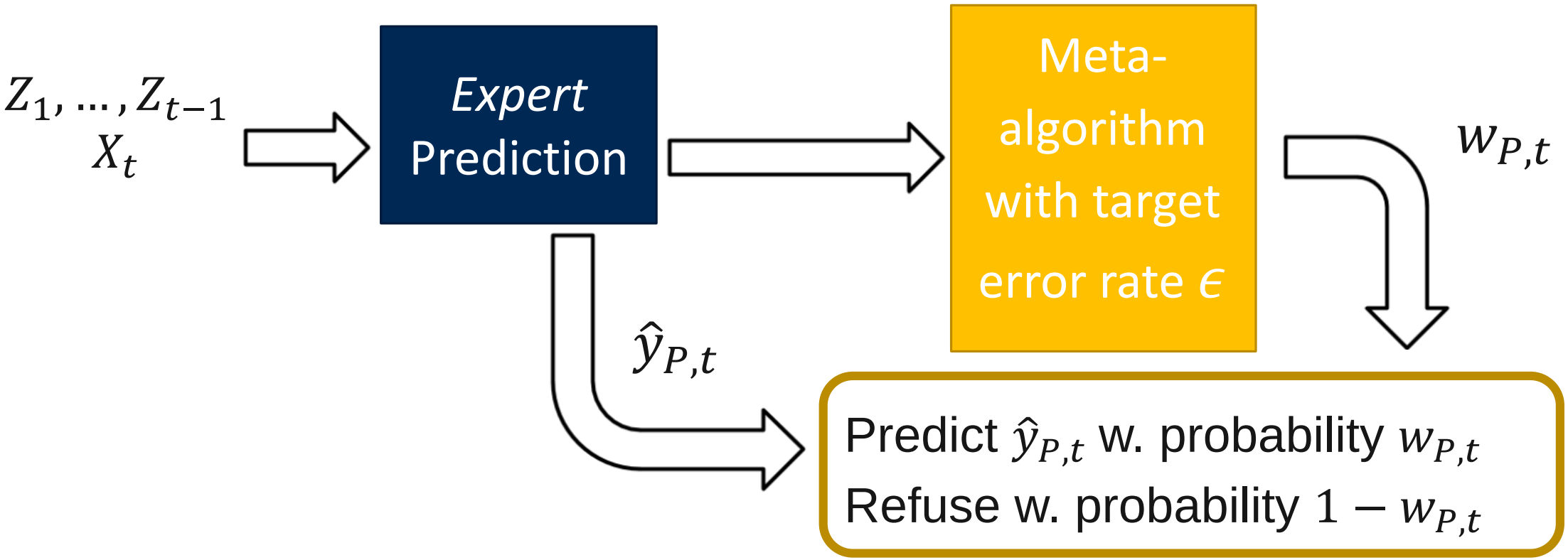
Crowd of experts with a refusing "expert"



Basic Idea: *SafePredict*, a meta-algorithm, takes predictions from experts and decides whether to pass them on to higher level application.

Refusing to predict is analogous to "gathering more data" before predicting

Online Prediction with Refusal



Prediction $\hat{y}_{P,t}$ or refusal $\hat{y}_D$ suffer a loss $l_{P,t}$, $l_D \in [0,1]$.

Mistakes can help us *learn* (i.e., $w_{P,t}$ changes with t depending on performance).

No assumptions on data or base predictor!

Nick Littlestone, and Manfred K. Warmuth, "The weighted majority algorithm." *Information and computation,* 1994.

Validity and Efficiency

Def. A meta-algorithm is ***valid*** if, as the number of predictions goes to ∞, the expected loss is less than a given target error rate.

Theorem 1.- With learning rate $\eta = \Theta(\frac{1}{\sqrt{V^*}})$, *SafePredict* is guaranteed *valid* for any P. Particularly $\frac{L_T^*}{T^*} - \epsilon = O\left(\frac{\sqrt{V^*}}{T^*}\right) = \left(\frac{1}{\sqrt{T^*}}\right)$.

Def. A meta-algorithm is ***efficient*** if, as the number of observed objects goes to ∞, the number of refusals is finite.

Theorem 2.- If $\limsup_{t\to\infty} \frac{L_{P,t}}{t} < \epsilon$ and $\eta T \to \infty$, then *SafePredict* is *efficient*.

Validity and Efficiency

Def. A meta-algorithm is ***valid*** if, as the number of predictions goes to ∞, the expected loss is less than a given target error rate.

Theorem 1.- Valid for any predictor if we pick a learning rate properly. Proof based on properly picking a regret bound

Def. A meta-algorithm is ***efficient*** if, as the number of observed objects goes to ∞, the number of refusals is finite.

Theorem 2.- Efficient if there exists a good enough predictor (we don't need to know it exists) and learning rate is *quick* enough.

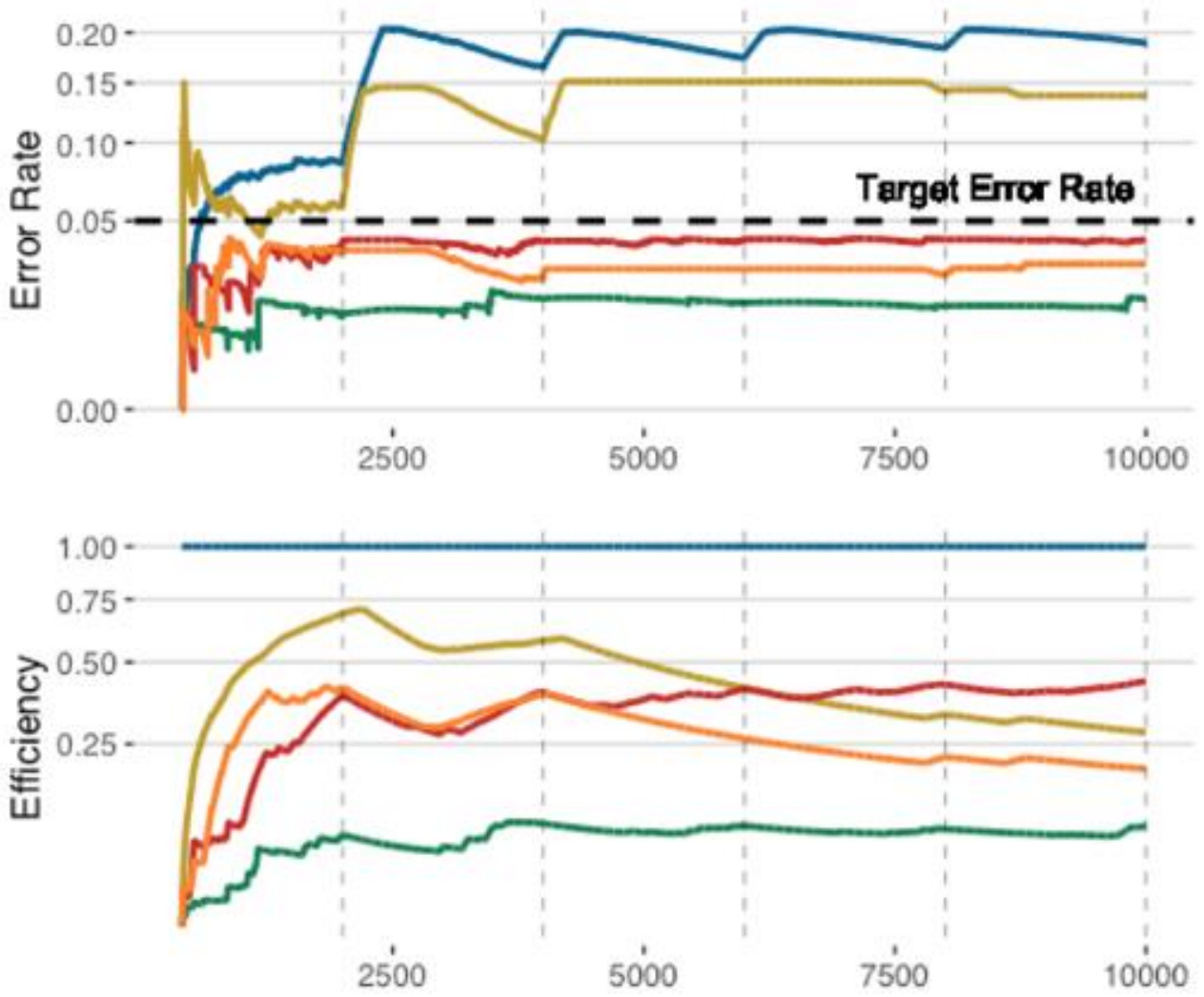
Experimental Results (MNIST)

Random label permutation every 2k points

- Change points are unknown a priori

Ideally, want a low error rate and high efficiency

- Base Predictor
- Confidence Based Refusal (CBR)
- *SafePredict*
- *SafePredict* + CBR
- Amnesic *SafePredict*
  (if 50% of the last 100 predictions were “Refuse”, forget history and reset weights)



Takeaway Message for *SafePredict*

- Works with *any* prediction algorithm (i.e., it's a meta-algorithm)
- Guarantees error rate for non-refused predictions
- **No** assumptions on data or predictor for guarantees
- In dynamic environments, amnesic heuristic can boost efficiency while remaining valid!
- Paper available on arXiv (with link to notebooks)