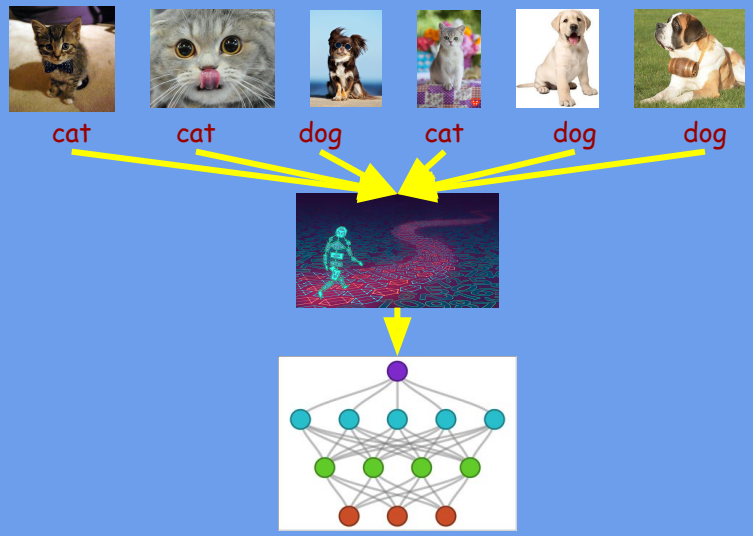
The Information-Theoretic Value of Unlabeled Data in Semi-Supervised Learning

Alexander Golovnev, Dávid Pál, Balázs Szörényi

IMCL 2019

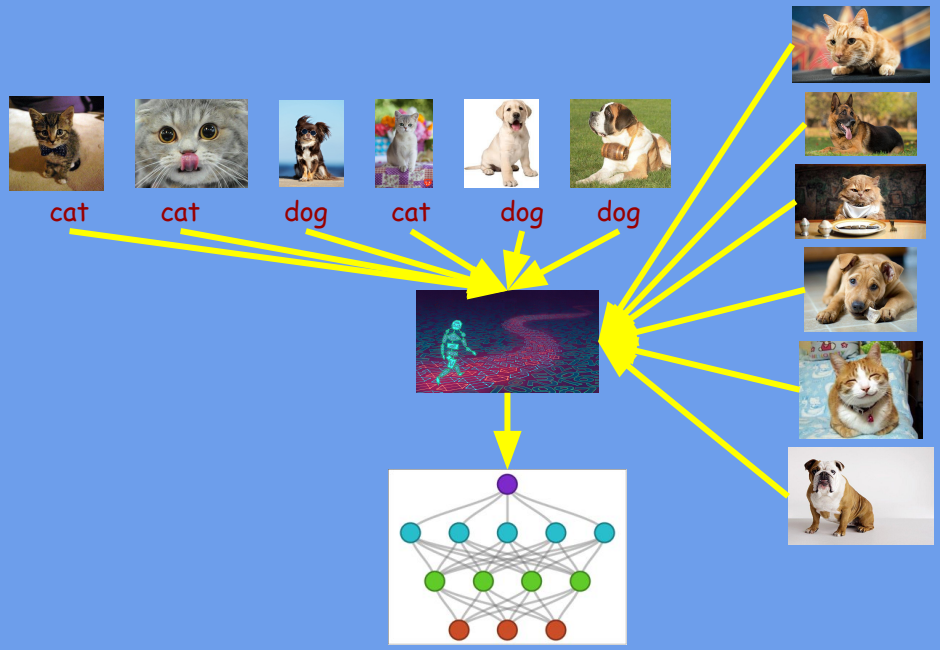
Do unlabeled data help?

Thought Experiment



A = Number of labeled examples required to guarantee 99% accuracy

Infinite amount of unlabeled data

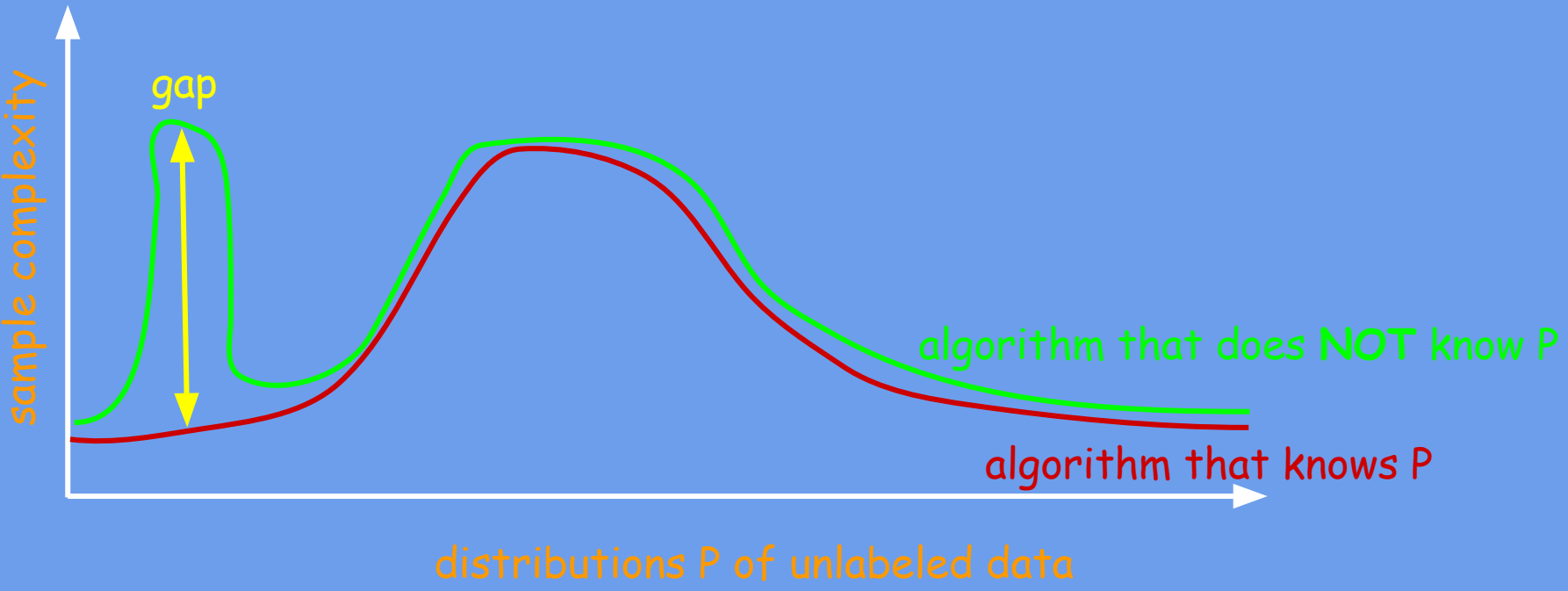


B = Number of labeled examples required to guarantee 99% accuracy

PAC model

- distribution P over a domain X
- class of binary classifiers H
- unknown target function f from H
- i.i.d. labeled sample $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m))$
- train classifier g
- $error(g) = Pr[f(x) \neq g(x)]$
- sample complexity = smallest m such that $err(g) < 0.01$ w.p. 95%

Sample complexity as a function of P



Theorem

For learning projections over $\{0,1\}^n$, the **multiplicative** gap of any algorithm that does **not** know P is $\Omega(\log(n))$.