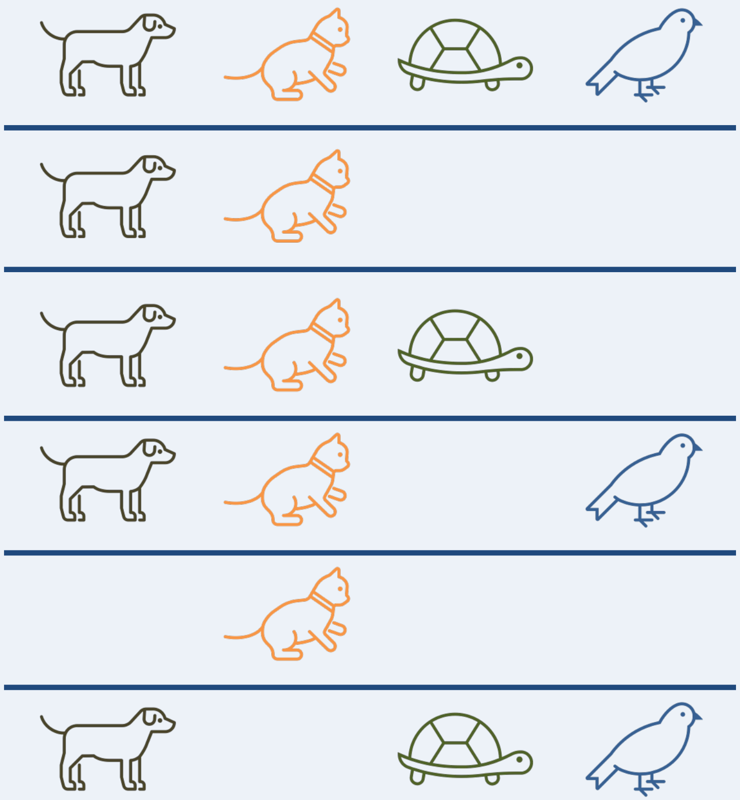
Subsampling Suffices for

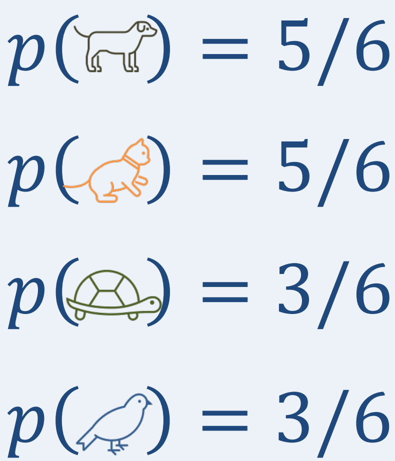
Adaptive Data Analysis

Focus today

Using Samples to understand the Population

What pets do people like?





Using Samples to understand the Population

Fact: With a sample of size

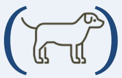
$$n \geq \Omega\left(\frac{\log q}{\varepsilon^2}\right)$$

q-many $p()$ queries will be within $\pm\varepsilon$ of their values in the overall population.

Proof: Union bound over q queries each with failure probability $\ll 1/q$.

For a single query, value of p_{sample} is mean of n independent $\mathrm{Ber}(p_{\text{population}})$. By Hoeffding's inequality,

$$\Pr\left[\left|p_{\text{sample}} - p_{\text{population}}\right| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 n}$$

$p($  $) = 5/6 \pm \varepsilon$

$p($  $) = 5/6 \pm \varepsilon$

$p($  $) = 3/6 \pm \varepsilon$

$p($  $) = 3/6 \pm \varepsilon$

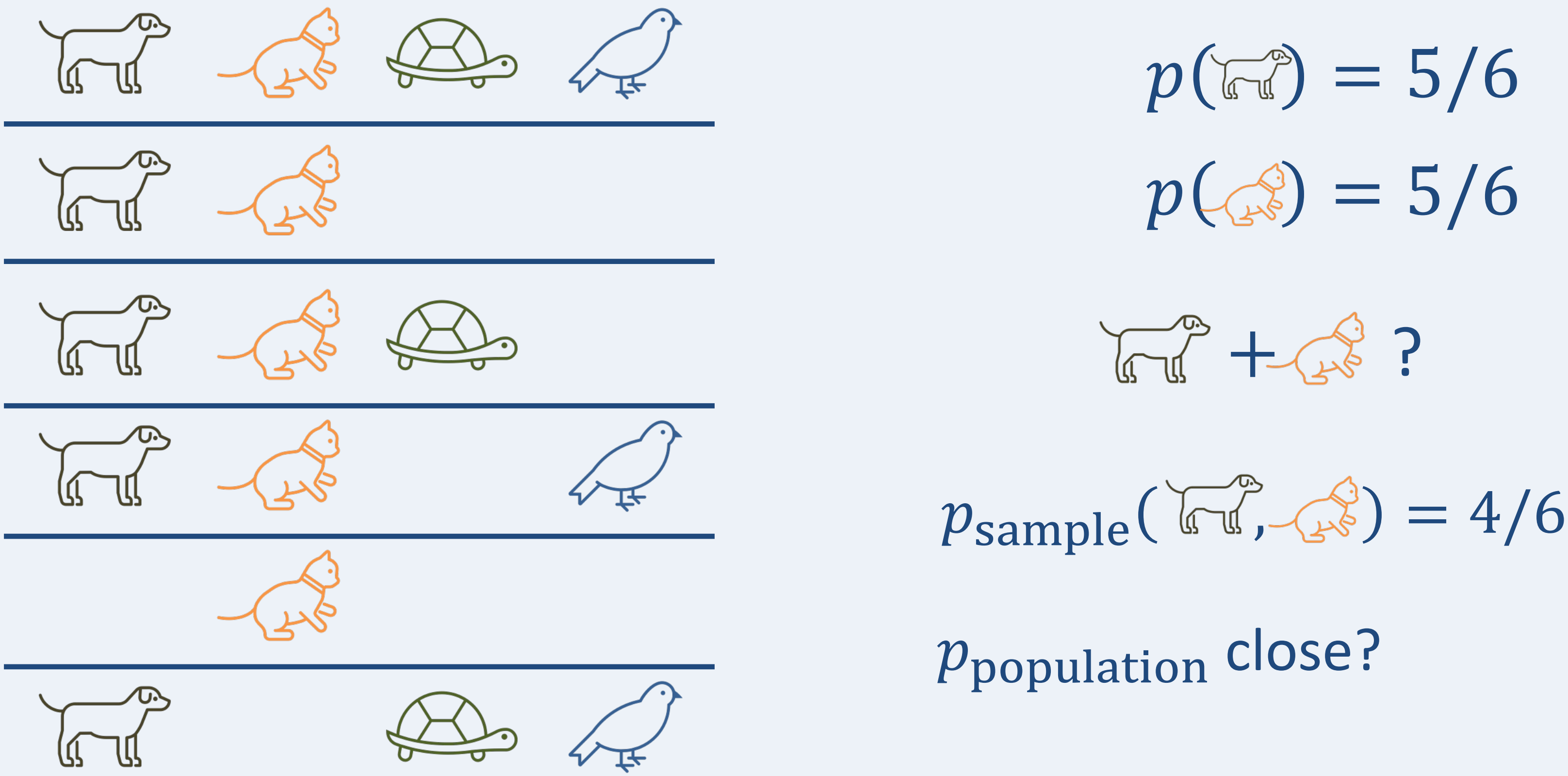
Using Samples to understand the Population

Other examples:

1. What fraction of patients on medication X experience remission?
2. What fraction of people will vote for candidate Y?
3. What fraction of concepts does a student understand?

Adaptive data analysis

What pets do people like?



Adaptive Data Analysis

How can we guarantee results are representative of the population even when the queries are chosen adaptively?

Proposed by [Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth 15]

Why adaptive data analysis is hard

Non-adaptive: Sample of size $n \geq \Omega\left(\frac{\log q}{\varepsilon^2}\right)$ suffices

Adaptive counterexample: Population distribution,

$$\mathcal{D} \coloneqq \text{Uniform}(\{1,2,\ldots,2n\})$$

Given sample $\boldsymbol{S} \sim \mathcal{D}^n$, for each $i \in \{1,2,\ldots,2n\}$), ask query:

$$y_i = p_{\boldsymbol{S}}(x \mapsto \mathbf{1}[x = i])$$

Track $T \coloneqq \{i \text{ where } y_i > 0\}$

After receiving response, ask query $x \mapsto \mathbf{1}[x \in T]$).

1. $p_{\boldsymbol{S}}(x \mapsto \mathbf{1}[x = i]) = 1$
2. $p_{\mathcal{D}}(x \mapsto \mathbf{1}[x = i]) \leq \frac{1}{2}$

Adaptive: With $q = 2n + 1$, can force error $\varepsilon \geq 1/2$.

Adaptive Data Analysis

How can we guarantee results are representative of the population even when the queries are chosen adaptively?

Proposed by [Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth 15]

Any ideas?

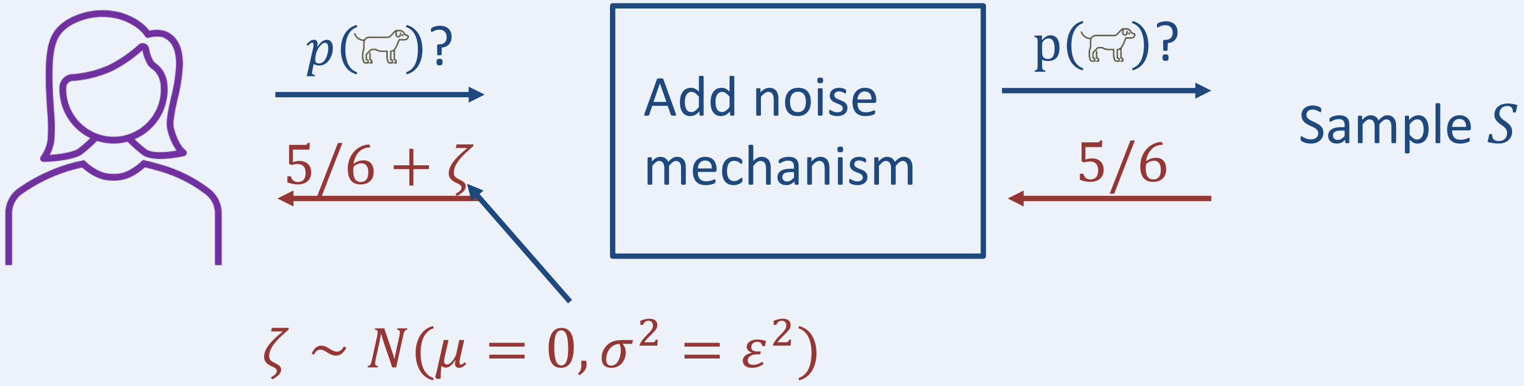
A simple solution

Take a fresh batch of $\approx 1/\varepsilon^2$ samples for each query.

Requires sample size of

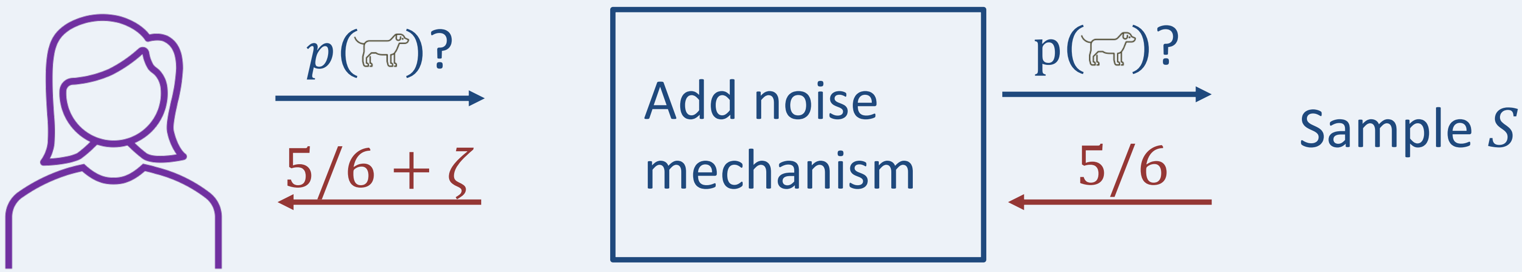
$$n \approx \frac{q}{\varepsilon^2}.$$

A better mechanism



Needs only $n = \tilde{O}\left(\frac{\sqrt{q}}{\varepsilon^2}\right)$ samples to answer q queries [DFHPR15, BNSSSU16].

Why is adding noise good?



Intuition: To ask a bad query,  must have lots of information about S.

Adding noise "hides" information about S. Formally, it ensures the query responses are differentially private.

Very cool active area of research on how to quantify private algorithms.

My research question

What minimal assumptions can we make about the queries to guarantee the results generalize, even without an explicit mechanism?

My solution: Sufficient for each query to take as input a random subsample and outputs few bits.

Subsampling queries

$\phi: X^w \to Y$

$\phi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_w)$

Sample $S \in X^n$

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_w$ chosen uniformly without replacement from S

Theorem (informal): If each $|Y|$ is small, results will be representative for q queries as long as the sample size satisfies

$$n \geq \Omega(w\sqrt{q}).$$

Compare to $n \geq wq$ required if we use a separate sample for each query.

Example application #1: Fraction queries

p(🐕)?

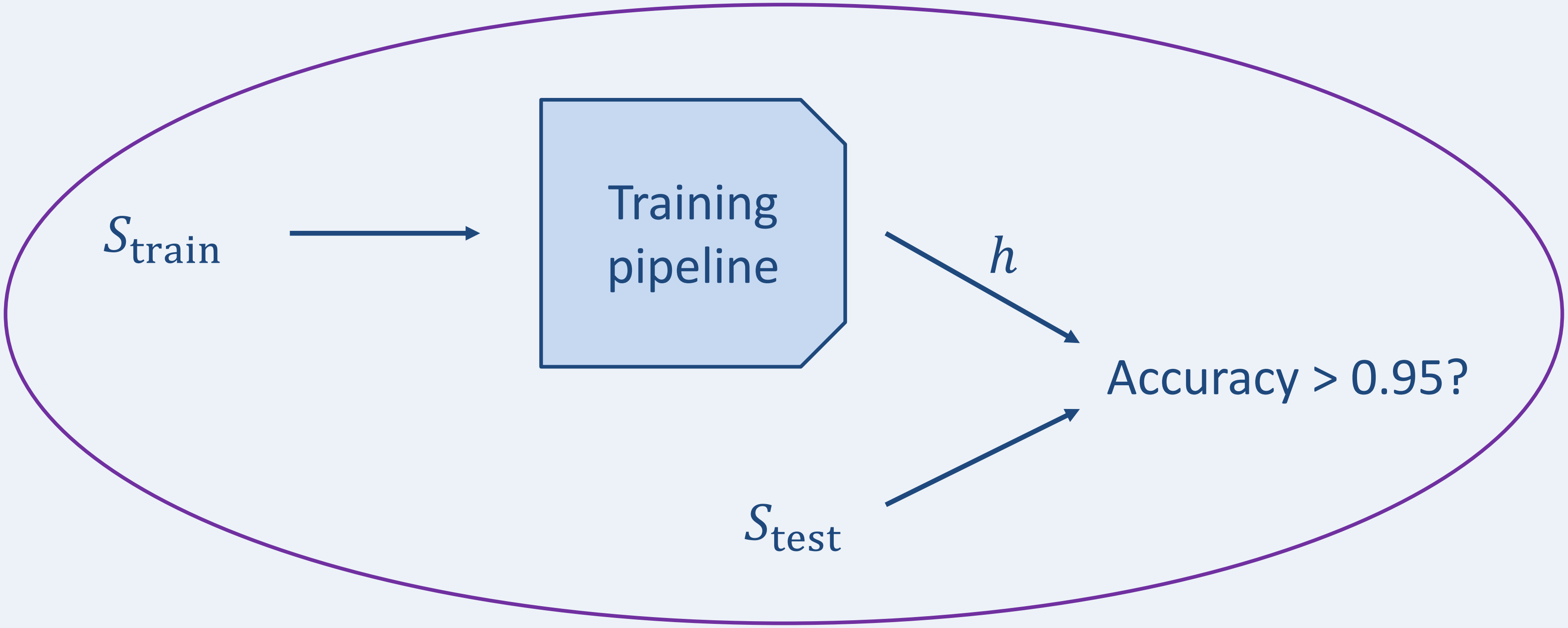
$[p($🐕$) \text{ in } (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_w)]$

Sample $S \in X^n$

$\boldsymbol{x}_i \sim \text{Unif}\,(S)$

This simple mechanism gives state of the art sample size-accuracy tradeoff.

Example application #2: Is a training pipeline accurate?



All one subsampling query with $w = |S_{\text{train}}| + |S_{\text{test}}|$ and $Y = \{0,1\}$

Questions?

$\phi: X^w \to Y$

$\phi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_w)$

Sample $S \in X^n$

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_w$ chosen uniformly without replacement from S

Theorem (informal): If each $|Y|$ is small, results will be representative for q queries if

$$n \geq \Omega(w\sqrt{q}).$$