More DP techniques

1. Composition, Advanced Composition
2. Sparse Vector
3. Blum-Ligett-Roth [BLR]:
   release a sanitized database
   that is DP, and accurate
   for a large family of queries
4. DP $\Rightarrow$ generalization

Sparse Vector Technique

Say we have a lot of queries but we only want to respond with numeric value if it is over a certain threshold, T

Example 1

Stream of data, people + their salaries. Output "above" threshold for people with salaries $> T$.

Example 2 Test model on holdout test set. Output "error" if error exceeds threshold T

Idea:

add noise to count and to query $f_i(D)$

When noisy $\hat{f}_i(D)$ exceeds noisy T,
output "exceeds threshold"
and otherwise output "⊥"

HALT when noisy count exceeds c

Algorithm Sparse $(D, \{f_i\}, T, c, \varepsilon, \delta)$

Let $G = \begin{cases} 2c/\varepsilon & \text{if } \delta = 0 \\ 32c \ln 1/\delta / \varepsilon & \text{otherwise} \end{cases}$

$\hat{T}_0 = T + Lap(G)$; count $= 0$

For $i = 1, 2, \ldots$

$v_i = Lap(2G)$

If $f_i(D) + v_i \geq \hat{T}_{count}$ then

output $a_i = \top$

count $=$ count $+ 1$

$\hat{T}_{count} = T + Lap(G)$

Else output $a_i = \bot$

If count $\geq c$ HALT

More DP techniques

1. Composition, Advanced Composition
2. Sparse Vector
3. Blum-Ligett-Roth [BLR]:
   release a sanitized database that is DP, and accurate for a large family Q of queries
4. DP $\Rightarrow$ generalization

BLR Mechanism

- We want to generate a (distribution over) synthetic databases that answers all queries from a family $\mathcal{Q}$ accurately.
- Use sample complexity bounds + Exponential Mechanism

BLR Mechanism

Let x be true database, and
y is candidate synthetic database

$$q(x,y) = -\max_{f \in Q} |f(x) - f(y)|$$

↖ utility of y wrt Q, x

Sample complexity Upper Bounds

Lemma $\forall x \in \mathbb{N}^{|\mathcal{X}|}$, and any collection of linear queries Q, there exists a database y of size $\log |Q| / \alpha^2$ such that

$$\max_{f \in Q} |f(x) - f(y)| \leq \alpha$$

BLR Mechanism

Let R = all db's $y \in \mathbb{N}^{|\mathcal{X}|}$, $\|y\|_1 = \log \frac{|Q|}{\alpha^2}$

Let $q(x, y) = -\max_{f \in Q} |f(x) - f(y)|$

Output $r \in R$ according to exponential mechanism $M_{exp}(x, q, R)$

BLR is $(\varepsilon, 0)$ private, and very accurate

More DP techniques

1. Composition, Advanced Composition
2. sparse Vector
3. Blum-Ligett-Roth [BLR]:
   release a sanitized database
   that is DP, and accurate
   for a large family of queries
4. DP $\Rightarrow$ generalization

DP $\Rightarrow$ generalization

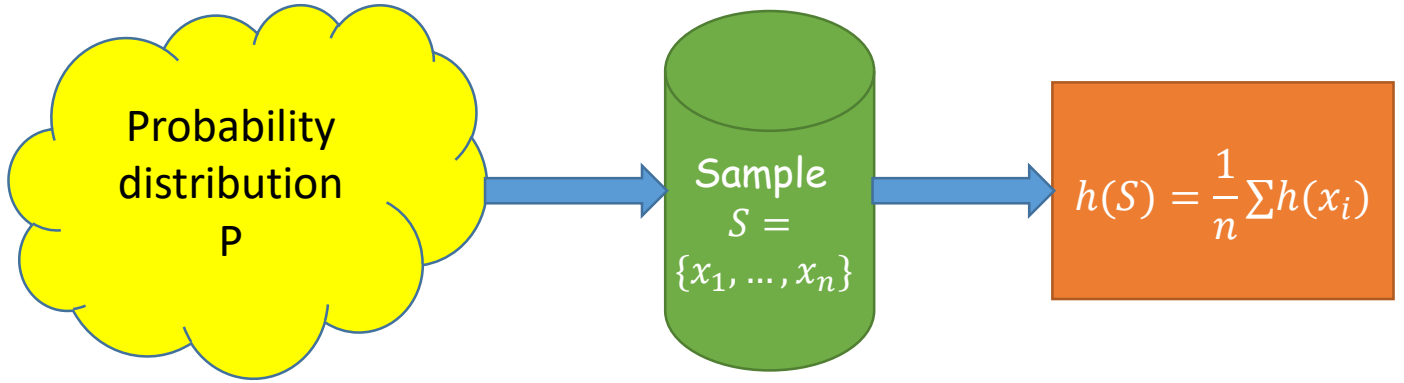
What is generalization?

Say we train a model on training set X,
where X = n labelled examples
ie. x_i : choose $u \sim \rho$, $x_i = (u, f(u))$

If model is accurate on X, then we
want to conclude model is accurate
on whole distribution

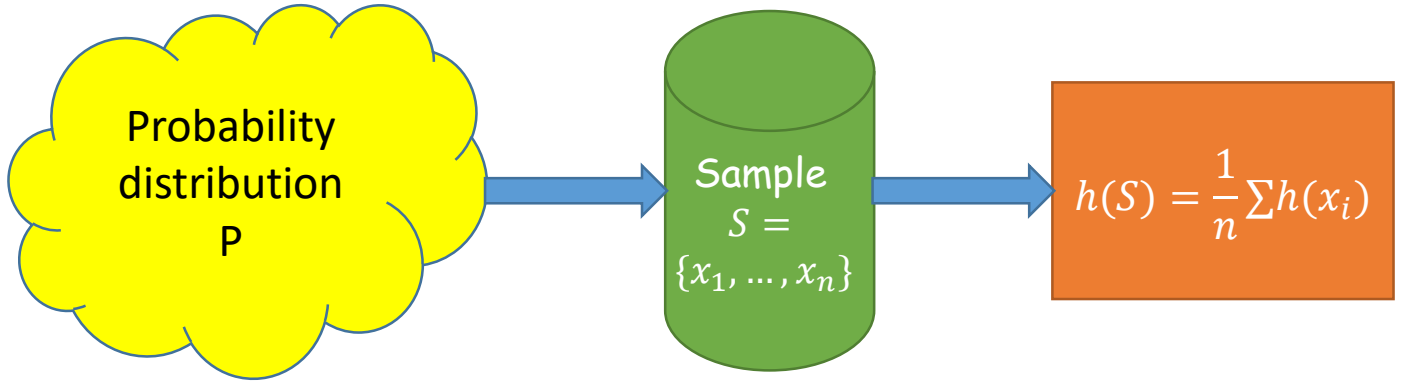
Intro: an estimation problem

- X: an arbitrary domain
- P: an (unknown) probability distribution over the domain X
- $h: X \to [0,1]$
- Estimate $h(P) = \underset{x \sim P}{\mathrm{E}}[h(x)]$
  - $S = \{x_1, \ldots, x_n\}$: a sample of n i.i.d. example drawn from P
  - Return $h(S) = \frac{1}{n}\sum h(x_i)$ as an estimation for $h(P)$
- How far is $h(S)$ from $h(P)$?
  - Intuitively, $h(S)$ estimates $h(P)$ well if n is large enough, but how large?



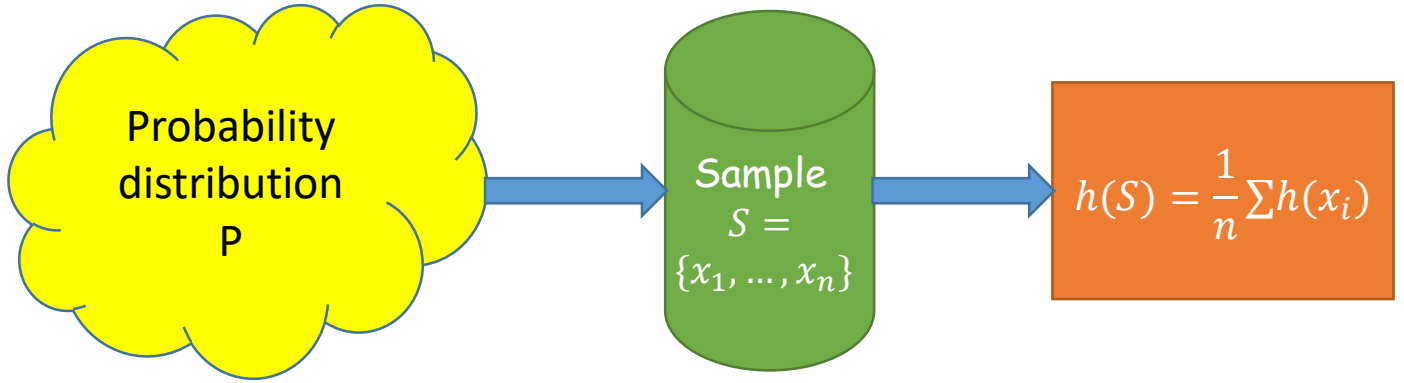
Intro: an estimation problem

- How large should the sample size n be?
- Tool: Hoeffding bound
  - $Z_1, \ldots, Z_n$: independent random variables
    - $Z_i \in [0,1]$ and $E[Z_i] = \mu$
  - $\hat{\mu} = \frac{1}{n}\sum z_i$
  - Theorem: for all $\alpha > 0$, $\Pr[\,|\hat{\mu} - \mu| \geq \alpha] \leq 2e^{-2n\alpha^2}$
- Using the Hoeffding bound:
  - $z_i = h(x_i)$
  - To get $|h(S) - h(P)| \leq \alpha$ with probability $\geq 1 - \beta$ suffices to take $n = O(\frac{\log\frac{1}{\beta}}{\alpha^2})$.



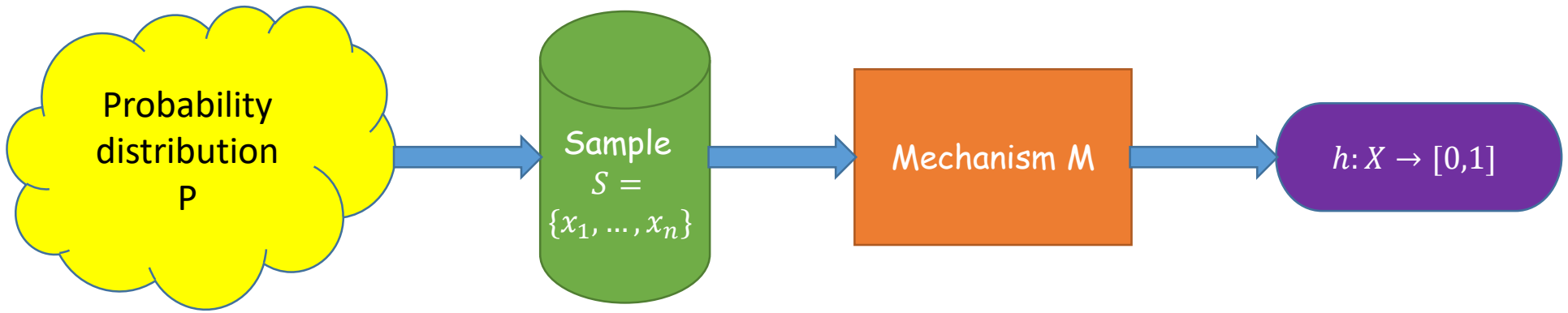
Simultaneously estimating a family of functions

- H: a family of functions $\{h: X \to [0,1]\}$
- Suffices to take $n = O(\frac{\log |\mathrm{H}| + \log\frac{1}{\beta}}{\alpha^2})$ samples to simultaneously estimate $h(P)$ within error α for all $h \in H$ with success probability $1 - \beta$
  - For each $h \in H$ we get (Hoeffding) $\Pr[|h(S) - h(P)| > \alpha] \leq 2e^{-2n\alpha^2} \leq \frac{\beta}{|H|}$.
  - Using union bound, $\Pr[\exists h \in H\ s.t. |h(S) - h(P)| > \alpha] \leq \beta$.



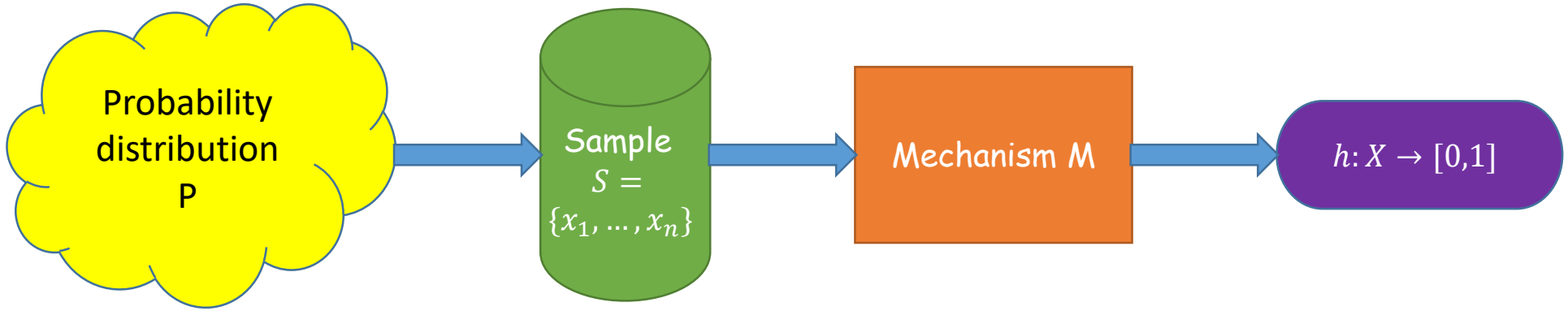
When h is chosen based on the sample

- Can't we use the Hoeffding bound?
- Let P be uniform over [0,1]
- Given $S = \{x_1, \dots, x_n\}$ let $\tilde{h}_S(x) = \begin{cases} 1 & if\ x \in S \\ 0 & otherwise \end{cases}$
- We get $\tilde{h}_S(S) = 1$ but $\tilde{h}_S(P) = 0$

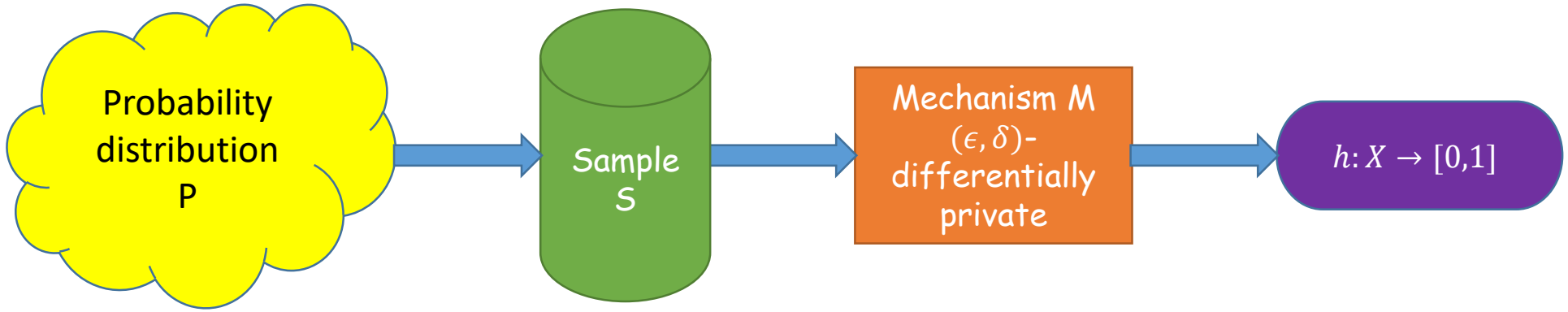


Generalization

- We say that a *hypothesis* $h: X \to [0,1]$ α-generalizes (w.r.t. S) if
$$|h(S) - h(P)| \leq \alpha$$
- What we saw:
  - When h is predetermined, $n = O(\frac{\log \beta}{\alpha^2})$ samples suffice for obtaining α-generalization
  - When H is predetermined, $n = O(\frac{\log |H| + \log \beta}{\alpha^2})$ samples suffice of simultaneously obtaining α-generalization for all H
  - Selection of h based on the sample can lead to "overfitting"



Differential privacy → generalization "on average"



- Intuition: "Overfitting is a common enemy"
- Theorem [McSherry, folklore]: $\left|\mathbb{E}[h(S)] - \mathbb{E}[h(P)]\right| \leq \epsilon + \delta$

Intuition: consider two experiments:

s_i : a random element of S

- $S = (s_1, \dots, s_n) \sim P$
- $z \sim P$
- $i \in_R [n]$
- $h \leftarrow M(S)$
- Return $h(s_i)$

$\approx$ **DP**

- $S = (s_1, \dots, s_n) \sim P$
- $z \sim P$
- $i \in_R [n]$
- $h \leftarrow M(S \setminus \{s_i\} \cup \{z\})$
- Return $h(s_i)$

s_i : a random element of P

Differential privacy → generalization "on average"

- Theorem: $\left|\mathbb{E}[h(S)] - \mathbb{E}[h(P)]\right| \leq 2\epsilon + \delta$
- Proof:

$$\mathbb{E}[h(S)] = \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)}[h(S)]$$

(reorder expectations)

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)}\ \mathop{\mathbb{E}}_{i\in_R[n]}[h(x_i)]$$

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{i\in_R[n]}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)}[h(x_i)]$$

(consider M' that takes output of M and applies it on x_i, then apply proposition)

$$\leq \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{i\in_R[n]}\left[e^{\epsilon}\ \mathop{\mathbb{E}}_{z\sim P;\, h\leftarrow M(S\setminus\{x_i\}\cup\{z\})}[h(x_i)] + \delta\right]$$

(rename z and x_i as $(S, z) \equiv (S \setminus \{x_i\} \cup \{z\}, x_i)$

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{i\in_R[n]}\left[e^{\epsilon}\ \mathop{\mathbb{E}}_{z\sim P;\, h\leftarrow M(S)}[h(z)] + \delta\right]$$

($\mathop{\mathbb{E}}_{z\sim P}[h(z)] = h(P)$)

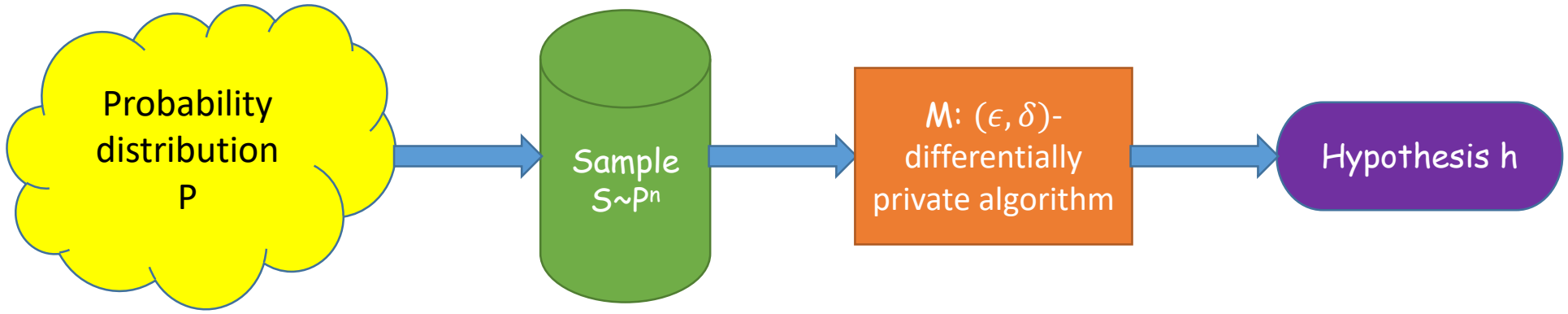
$$= \mathrm{e}^{\epsilon}\ \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)} h(P) + \delta$$

($e^{\epsilon} \leq 1 + 2\epsilon$ for $\epsilon < 1$)

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)} h(P) + 2\epsilon + \delta$$

(for other direction: let $h'(x) = 1 - h(x)$)

Differential privacy → generalization (summary)



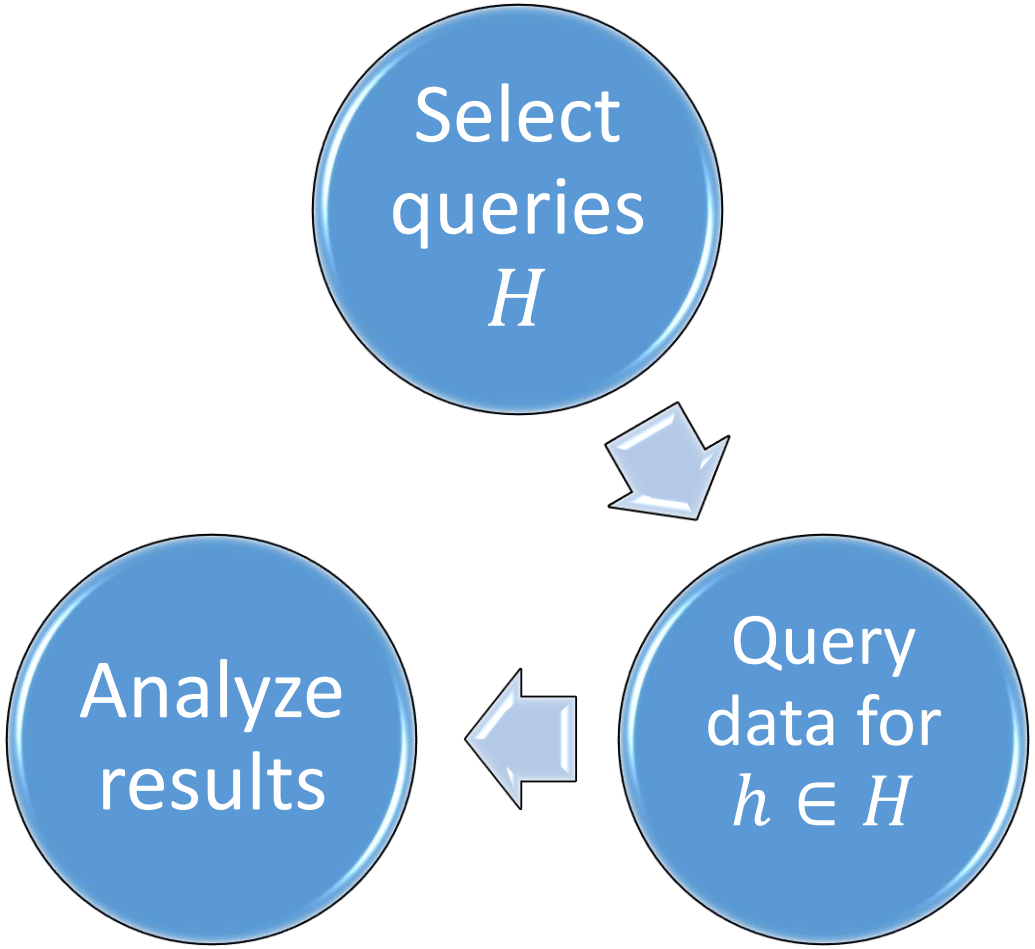
- Define: $h(S) = \frac{1}{n}\sum h(s_i)$ and $h(P) = \Pr_{s\sim P}[h(s)]$

Theorem [McSherry, folklore]:	$\mathbb{E}_{\substack{S\sim P \\ h\leftarrow M(S)}}[h(S)] \approx \mathbb{E}_{\substack{S\sim P \\ h\leftarrow M(S)}}[h(P)]$	Expectation
Theorem [DFHPRR'15]:	$\Pr_{\substack{S\sim P \\ h\leftarrow M(S)}}[\lvert h(S)-h(P)\rvert > \epsilon] \leq \delta^{\epsilon}$	High probability
Tight theorem [BNSSSU'16] $(n \geq O(\frac{\ln\frac{1}{\delta}}{\epsilon^2}))$:	$\Pr_{\substack{S\sim P \\ h\leftarrow M(S)}}[\lvert h(S)-h(P)\rvert > \epsilon] \leq \delta/\epsilon$	High probability

- Theorem: Let M be (ϵ, δ)-differentially private. Let S be $n \geq O(\frac{\ln\frac{1}{\delta}}{\epsilon^2})$ i.i.d. samples from an underlying distribution P. Interpret M(S) as a hypothesis h. Then,

$$\Pr[|E_S(h) - E_P(h)| > \epsilon] \leq O(\frac{\delta}{\epsilon})$$

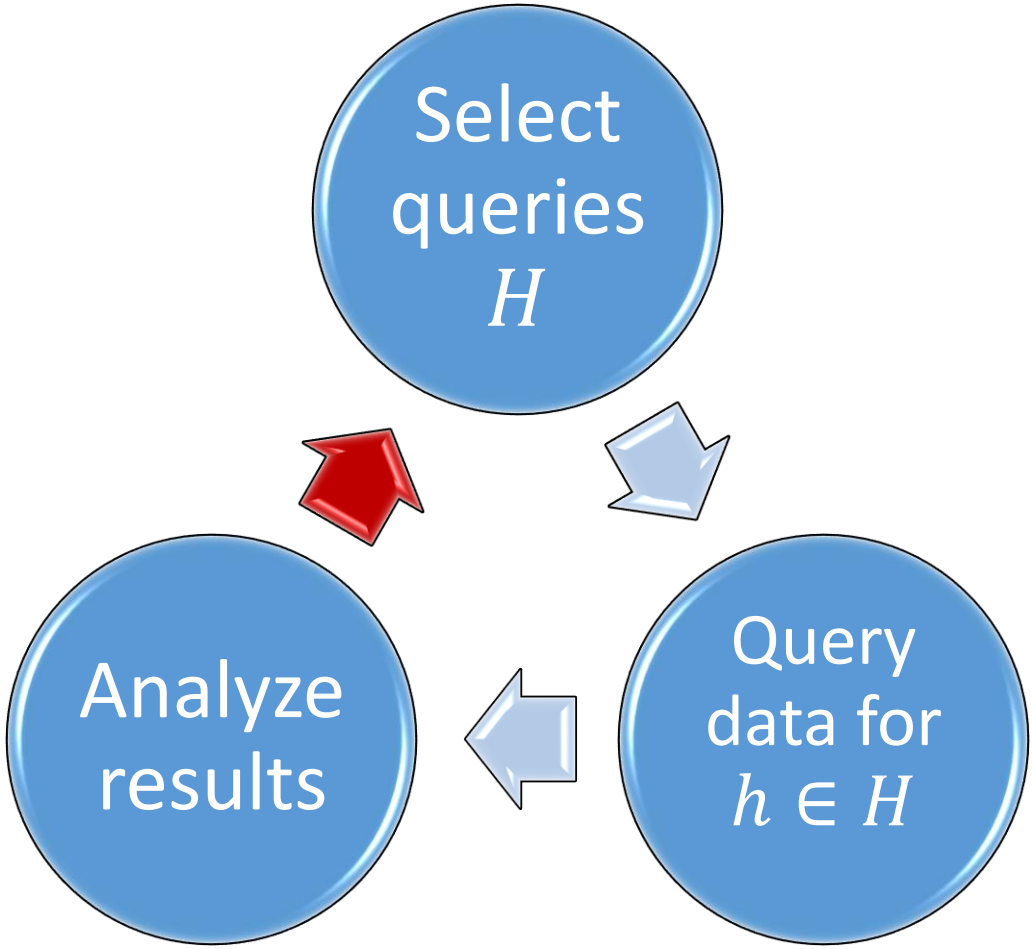
In theory* …



Statistically valid if sample size large enough ($\approx \log |H|$)

*In theory, theory and practice are the same. In practice, they are not. [A. Einstein]

In practice*



Analysts makes adaptive decisions:

- Queries selected based on the results of previous analyses
- Risk of false discoveries!
- A real problem. Lots of published research results are wrong!
- Almost all existing approaches to ensuring generalization assume the entire data-analysis procedure is fixed ahead of time

*In theory, theory and practice are the same. In practice, they are not. [A. Einstein]

Application to adaptive querying

- Differential privacy closed under post processing
  - Robust generalization: further post-processing unlikely to generate a non-generalizing hypothesis!
  - In standard learning, a model (that generalizes) may inadvertently reveal the sample, and hence lead to a non-generalizing hypothesis!
- Differential privacy closed under adaptive composition
  - [DFHPRR'15]: Even adaptive querying with differential privacy would not lead to a non-generalizing hypothesis

SUMMARY

Many DP mechanisms that can be mixed & matched:

- Laplace, Gaussian
- Sparse Vector
- Subsampling
- Advanced composition
- Exponential Mechanism

SUMMARY

Many DP mechanisms that can be mixed & matched:

- Laplace, Gaussian
- Sparse Vector
- Subsampling
- Advanced composition
- Exponential Mechanism

DP connected to generalization in ML and to hypothesis testing

OTHER DIRECTIONS (not discussed)

- Pan privacy: entire state of computation at all points in time is DP
- Local privacy / privacy in distributed systems
- Privacy under continual observation

SUMMARY

Many DP mechanisms that can be mixed & matched:

- Laplace, Gaussian
- Sparse Vector
- Subsampling
- Advanced composition
- Exponential Mechanism

DP connected to generalization in ML and to hypothesis testing