Generalization and Privacy

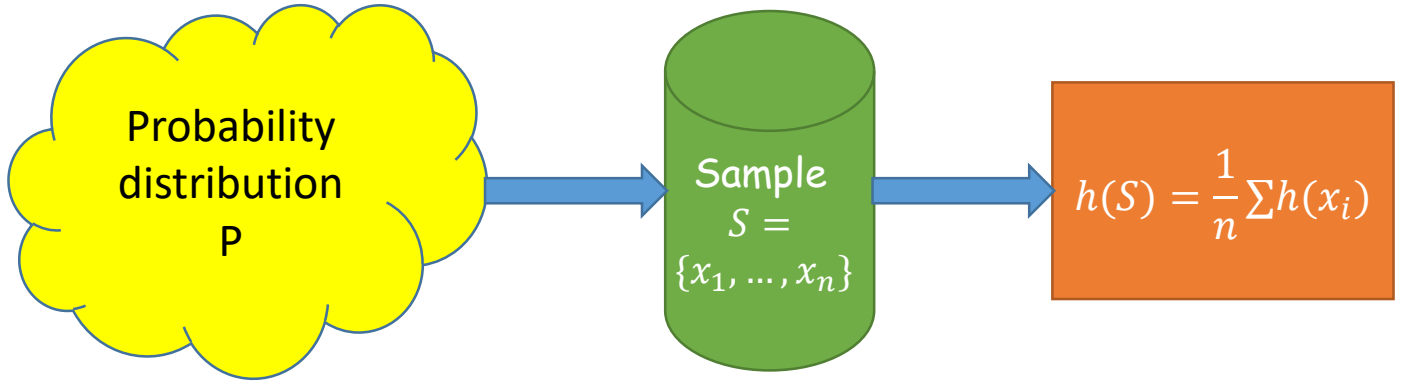
Kobbi Nissim, Georgetown University

Bar-Ilan Winter School on Differential Privacy

February 2017

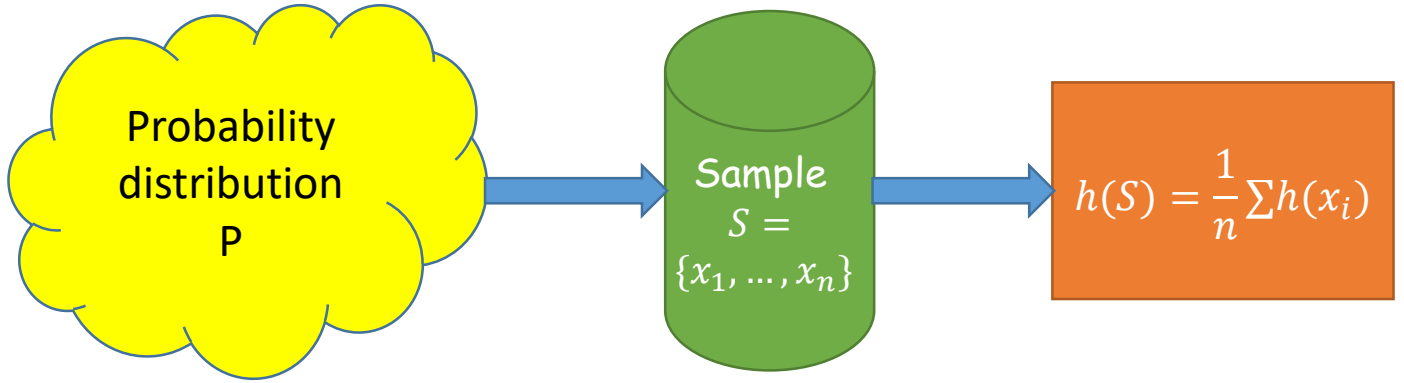
Intro: an estimation problem

- X: an arbitrary domain
- P: an (unknown) probability distribution over the domain X
- $h: X \to [0,1]$
- Estimate $h(P) = \mathop{\mathrm{E}}_{x \sim P}[h(x)]$
  - $S = \{x_1, \ldots, x_n\}$: 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?



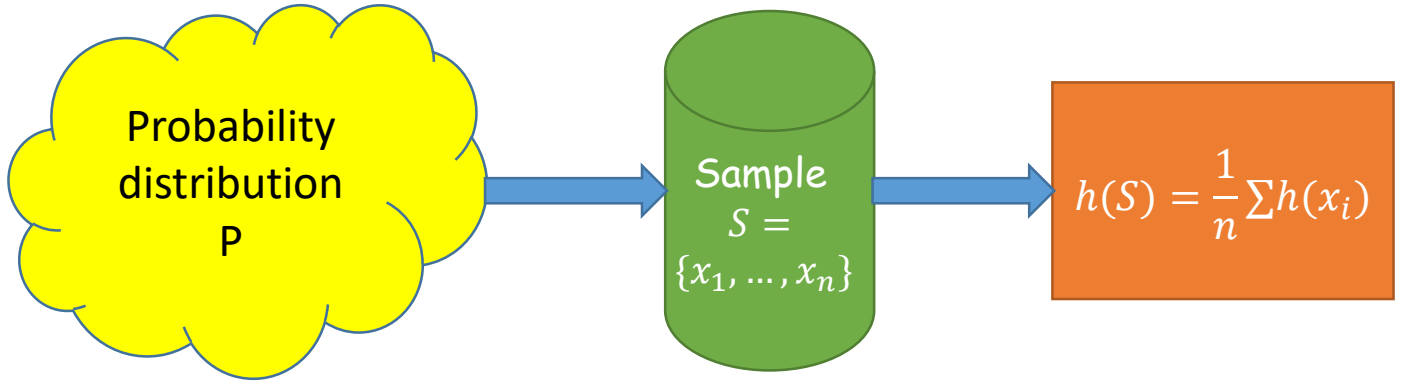
Intro: an estimation problem

- 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 $n = O(\frac{\log\frac{1}{\beta}}{\alpha^2})$.



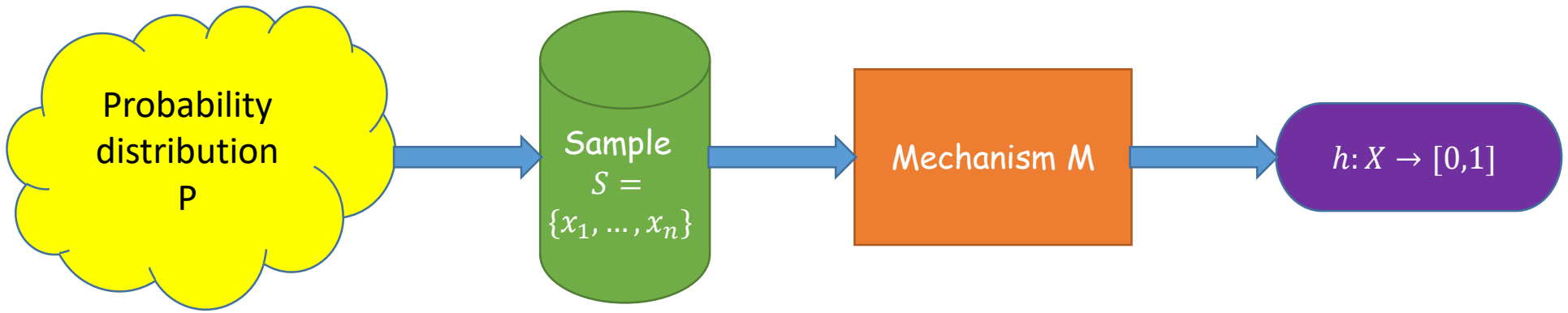
Simultaneously estimating a family of functions

- 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, $\Pr[\exists h \in H\ s.t. |h(S) - h(P)| > \alpha] \leq \beta$.



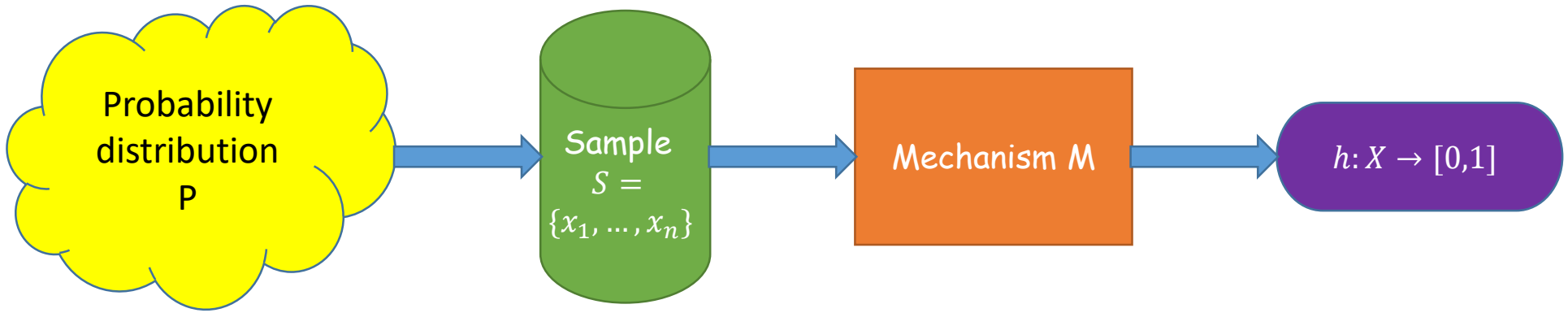
When h is chosen based on the sample

- 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 $\tilde{h}_S(P) = 0$

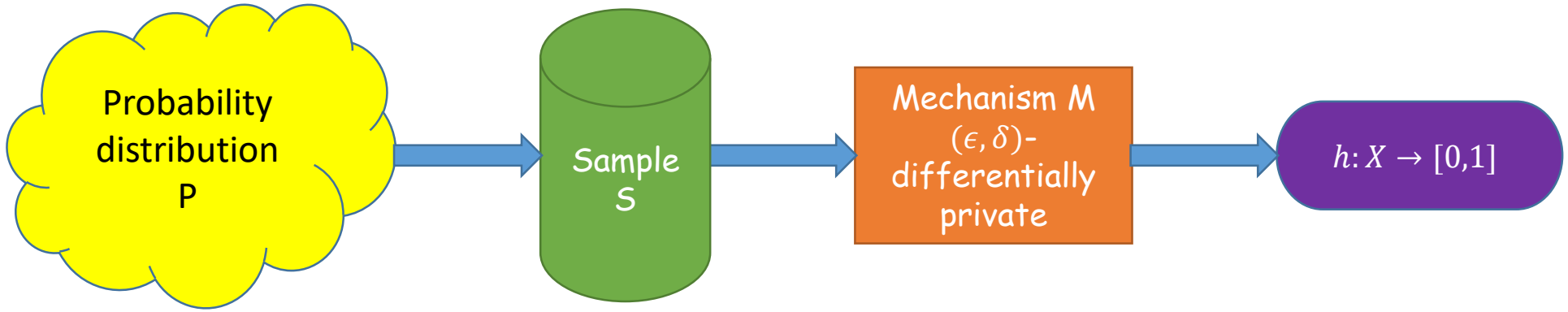


Generalization

- 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"

- A simple proposition:
  - $M: X^n \to [0,1]$: (ϵ, δ)-differentially private
  - S, S' neighboring datasets
  - Then $\mathop{\mathbb{E}}_{rand\ of\ M}[M(S)] \leq e^{\epsilon} \mathop{\mathbb{E}}_{rand\ of\ M}[M(S')] + \delta$
- Proof:

$$\mathop{\mathbb{E}}_{rand\ of\ M}[M(S)] = \int_0^1 Pr[M(S) > t]\, dt$$

$$\mathbb{E}[\mathrm{Y}] = \int_0^1 \Pr[Y > t]\, dt$$ for r.v. Y taking values in [0,1]

$$\leq \int_0^1 [\mathrm{e}^{\epsilon} Pr[M(S') > t] + \delta]\, dt$$ (differential privacy)

$$= e^{\epsilon} \mathop{\mathbb{E}}_{rand\ of\ M}[M(S')] + \delta$$

Differential privacy → generalization "on average"

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

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

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

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

- Proof strategy:

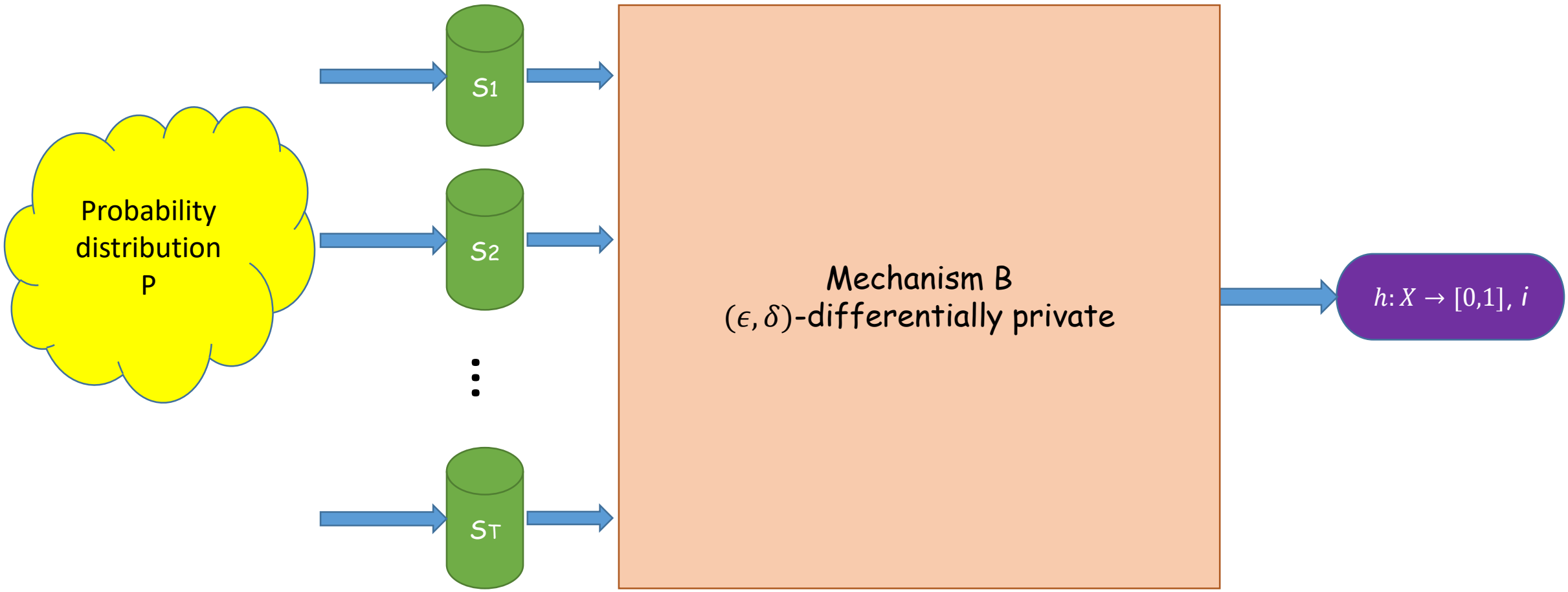
Begin with folklore guarantee (in expectation) M expects a single dataset $S = \{s_1, \dots, s_n\}$ and outputs a predicate h	$\left\lvert \mathop{\mathbb{E}}_{\substack{S \sim P \\ h \leftarrow M(S)}} [h(S)] - \mathop{\mathbb{E}}_{\substack{S \sim P \\ h \leftarrow M(S)}} [h(P)] \right\rvert \leq \epsilon + \delta$
Amplify to obtain high probability guarantee ($n \geq O(\frac{\ln \frac{1}{\delta}}{\epsilon^2})$)	$\mathop{\Pr}_{\substack{S \sim P \\ h \leftarrow M(S)}} [\lvert h(S) - h(P) \rvert > \epsilon] \leq \delta / \epsilon$

Differential privacy → generalization w.h.p.

- Proof strategy:

Begin with folklore guarantee (in expectation) M expects a single dataset $S = \{s_1, \dots, s_n\}$ and outputs a predicate h	$\left\lvert \mathop{\mathbb{E}}_{\substack{S \sim P \\ h \leftarrow M(S)}} [h(S)] - \mathop{\mathbb{E}}_{\substack{S \sim P \\ h \leftarrow M(S)}} [h(P)] \right\rvert \leq \epsilon + \delta$
Modify folklore guarantee to enable amplification B expects T sub-datasets $\vec{S} = (S_1, S_2, \dots, S_T)$ and outputs an index t of a sub-dataset and a predicate h	$\left\lvert \mathop{\mathbb{E}}_{\substack{\vec{\boldsymbol{S}} \sim \boldsymbol{P} \\ (\boldsymbol{h},\boldsymbol{t}) \leftarrow \mathrm{B}(\vec{\boldsymbol{S}})}} [\boldsymbol{h}(\boldsymbol{S_t})] - \mathop{\mathbb{E}}_{\substack{\vec{\boldsymbol{S}} \sim \boldsymbol{P} \\ (\boldsymbol{h},\boldsymbol{t}) \leftarrow \mathrm{B}(\vec{\boldsymbol{S}})}} [\boldsymbol{h}(\boldsymbol{P})] \right\rvert \leq \epsilon + T\delta$
Amplify to obtain high probability guarantee ($n \geq O(\frac{\ln \frac{1}{\delta}}{\epsilon^2})$)	$\Pr_{\substack{S \sim P \\ h \leftarrow M(S)}} [\lvert h(S) - h(P) \rvert > \epsilon] \leq \delta/\epsilon$

Modified folklore guarantee - setup



Proof of the modified folk guarantee

$\mathcal{B}$ = private alg. with

Input: T *sub-databases* $\vec{S} = (S_1, S_2, \dots, S_T)$ iid from P

Output: Predicate h and index $1 \le t \le T$

- Notation: $S_t = (x_{t,1}, \dots, x_{t,n})$

$$\mathop{\mathbb{E}}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}[h(S_t)] = \sum_{m=1}^{T} \mathop{\mathbb{E}}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}\left[1_{\{t=m\}} \cdot h(S_m)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \mathop{\mathbb{E}}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}\left[1_{\{t=m\}} \cdot h(x_{m,i})\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \mathop{\Pr}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}\left[1_{\{t=m\}} \cdot h(x_{m,i}) = 1\right]$$

- Given $\vec{S}$, (m, i), z define $\vec{S}^{(x_{m,i}:z)}$ to be as $\vec{S}$ after replacing $x_{m,i}$ with z

$$\le \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T}\left(e^{\epsilon} \mathop{\Pr}_{\substack{z,\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}\left(\vec{S}^{(x_{m,i}:z)}\right)}}\left[1_{\{t=m\}} \cdot h(x_{m,i}) = 1\right] + \delta\right)$$

Proof of the modified folk guarantee

- Given $\vec{S}, (m, i), z$ define $\vec{S}^{(x_{m,i}:z)}$ to be as $\vec{S}$ after replacing $x_{m,i}$ with z

$$\mathop{\mathbb{E}}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}[h(S_t)] \le \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T}\left(e^{\epsilon}\mathop{\Pr}_{\substack{z,\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}\left(\vec{S}^{(x_{m,i}:z)}\right)}}\left[1_{\{t=m\}}\cdot h(x_{m,i})=1\right]+\delta\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T}\left(e^{\epsilon}\mathop{\mathbb{E}}_{\substack{z,\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}\left(\vec{S}^{(x_{m,i}:z)}\right)}}\left[1_{\{t=m\}}\cdot h(x_{m,i})\right]+\delta\right)$$

- Every $\vec{S}^{(x_{m,i}:z)}$ above contains iid samples from P
- $x_{m,i}$ is independent of $\vec{S}^{(x_{m,i}:z)}$

$$=\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T}\left(e^{\epsilon}\mathop{\mathbb{E}}_{\substack{z,\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}\left[1_{\{t=m\}}\cdot h(z)\right]+\delta\right)=e^{\epsilon}\mathop{\mathbb{E}}_{\substack{z,\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}[h(z)]+T\delta$$

$$=e^{\epsilon}\mathop{\mathbb{E}}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}[h(P)]+T\delta\le\mathop{\mathbb{E}}_{\substack{\vec{S}\sim P \\ (h,t)\leftarrow\mathcal{B}(\vec{S})}}[h(P)]+2\epsilon+T\delta$$

Concluding the proof:

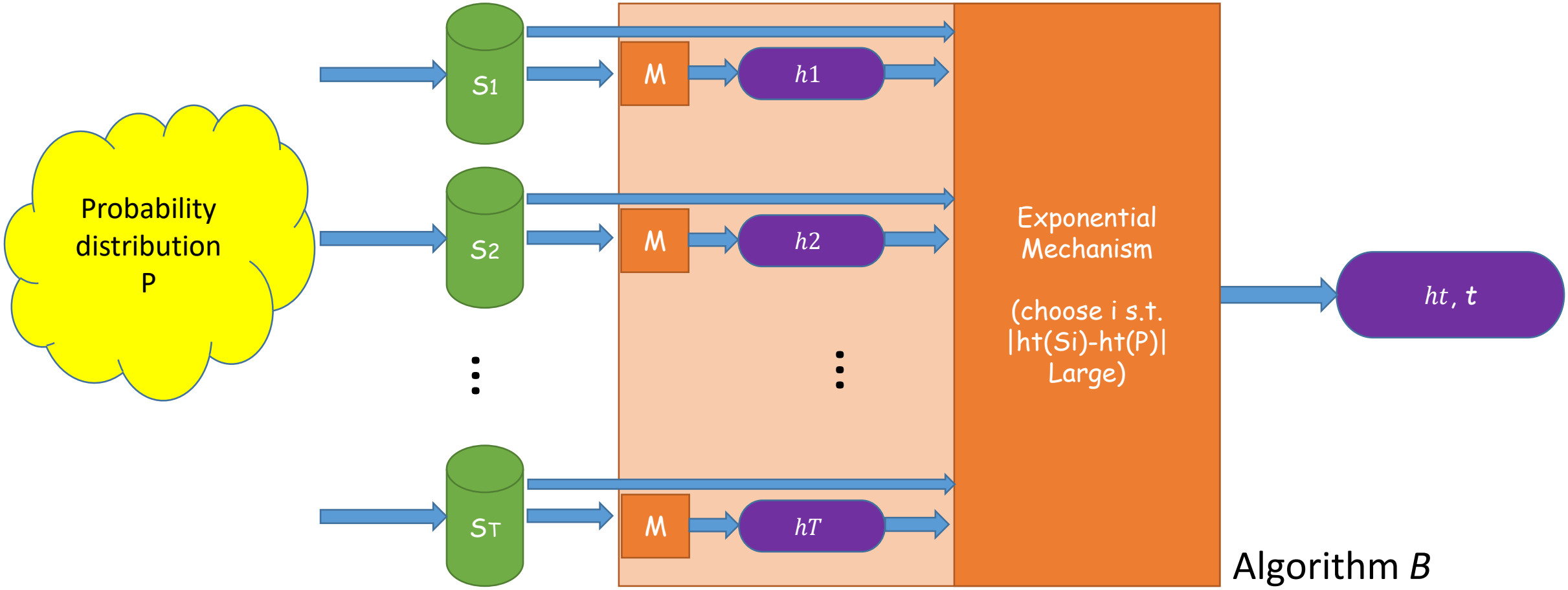
Modified folklore guarantee

B expects T sub-datasets $\vec{S} = (S_1, S_2, \dots, S_T)$ and outputs an index t of a sub-dataset and a predicate h

$$\left| \mathop{\mathbb{E}}_{\substack{\vec{S} \sim P \\ (h,t) \leftarrow \mathrm{B}(\vec{S})}} [h(S_t)] - \mathop{\mathbb{E}}_{\substack{\vec{S} \sim P \\ (h,t) \leftarrow \mathrm{B}(\vec{S})}} [h(P)] \right| \leq \epsilon + T\delta$$

- Given (ϵ, δ)-DP M and distribution P s.t. M outputs h with large $|h(S) - (P)|$ w.p. $\geq \delta/\epsilon$
- Create (ϵ, δ)-DP B that expects $T \approx \epsilon/\delta$ sub-datasets:
  - B executes M on each of the T sub-datasets to get predicates $h_1, \dots, h_T$
  - W.h.p. there exists t such that $|h_t(S_t) - h_t(P)|$ is large
  - B identifies (with DP) such a sub-dataset t and outputs t, h_t
  - W.h.p. $|h(S_t) - h_t(P)|$ is large, contradicting the modified theorem! M cannot exist!

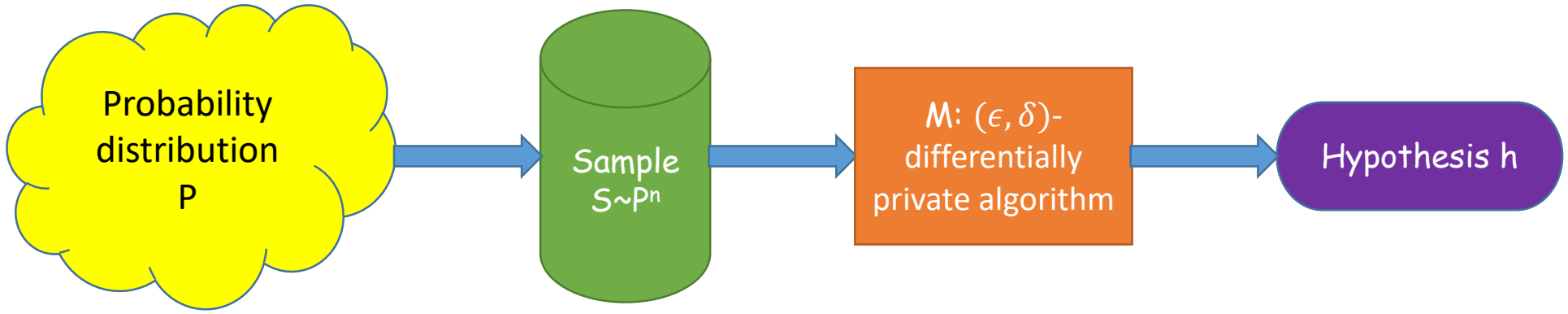
Modified folklore guarantee – proof construction



Notes:

Proof by contradiction: Algorithm B only used in proof, M does not need to be modified
Algorithm B ”knows” the underlying distribution P

Differential privacy → generalization (summary)



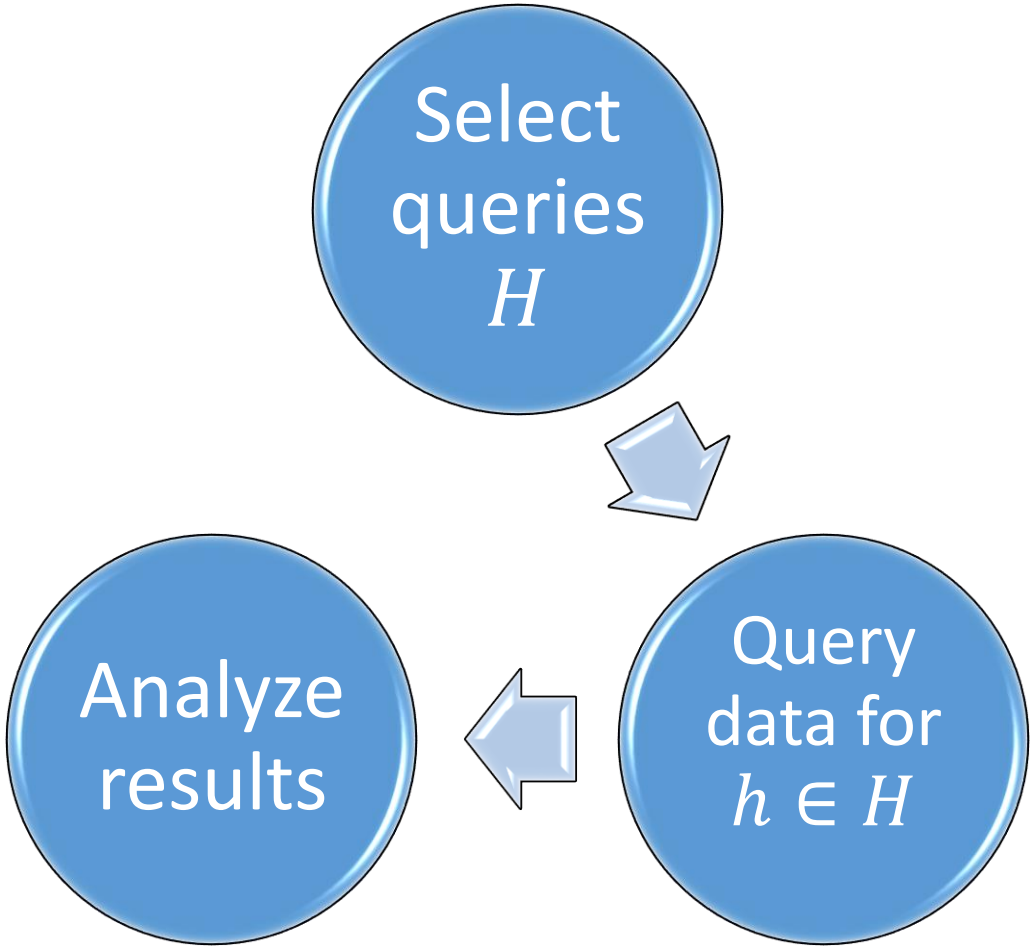
- Define: $h(S) = \frac{1}{n}\sum h(s_i)$ and $h(P) = \Pr_{s\sim P}[h(s)]$

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] $(n \geq O(\frac{\ln\frac{1}{\delta}}{\epsilon^2}))$:	$\Pr_{\substack{S\sim P \\ h\leftarrow M(S)}}[\lvert h(S)-h(P)\rvert > \epsilon] \leq \delta/\epsilon$	

- 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,

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

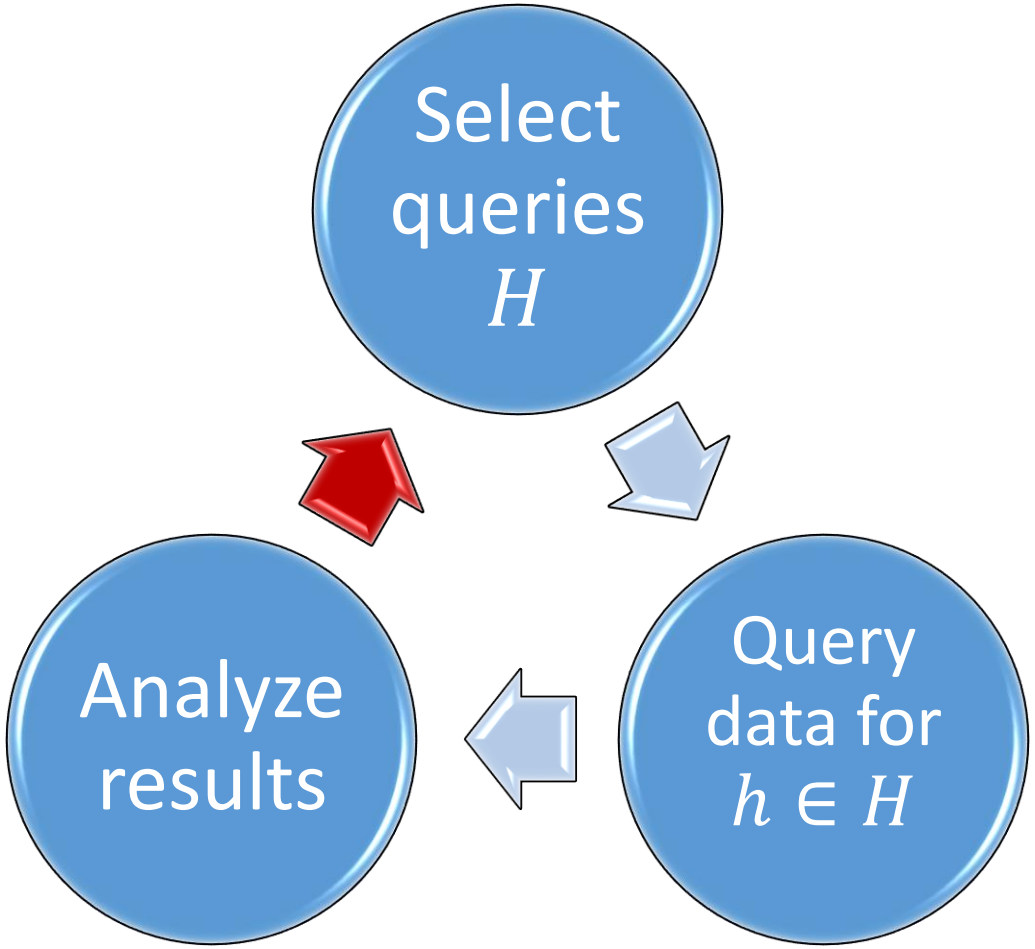
In theory* …



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

Application to adaptive querying

- Can import tools developed for answering queries adaptively with differential privacy!
- In particular, differential privacy allows approximating $h(S) = \frac{1}{n}\sum h(s_i)$ for $k \approx n^2$ adaptively selected predicates $h_1, \dots, h_k$

Upper bounds:

- [DFHPRR'15]: Efficient mechanism that w.h.p. answers any k adaptively chosen queries $h_1, \dots, h_k$ within accuracy α given $n = \tilde{O}\left(\sqrt{k}/\alpha^{2.5}\right)$ samples
- [BNSSSU]: Sample complexity reduced to $n = \tilde{O}\left(\sqrt{k}/\alpha^{2}\right)$

Lower bound:

- [Hardt Ullman 14, Steinke Ullman 15]: Any efficient mechanism that answers k adaptive queries within accuracy α requires $n = \Omega\left(\sqrt{k}/\alpha\right)$

References

- Moritz Hardt, Jonathan Ullman: **Preventing False Discovery in Interactive Data Analysis Is Hard.** FOCS 2014
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, Aaron Leon Roth: **Preserving Statistical Validity in Adaptive Data Analysis.** STOC 2015
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, Aaron Roth: **Generalization in Adaptive Data Analysis and Holdout Reuse.** NIPS 2015
- Thomas Steinke, Jonathan Ullman: **Interactive Fingerprinting Codes and the Hardness of Preventing False Discovery.** COLT 2015
- Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, Jonathan Ullman: **Algorithmic stability for adaptive data analysis.** STOC 2016
- Thomas Steinke, Jonathan Ullman: **Subgaussian Tail Bounds via Stability Arguments.** 2017