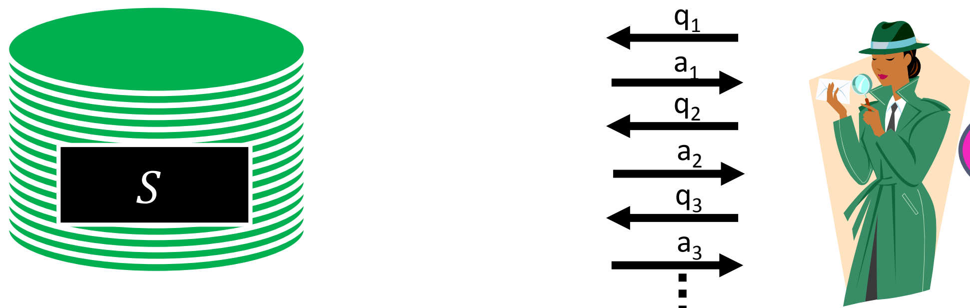
Universally Adaptive Data Analysis

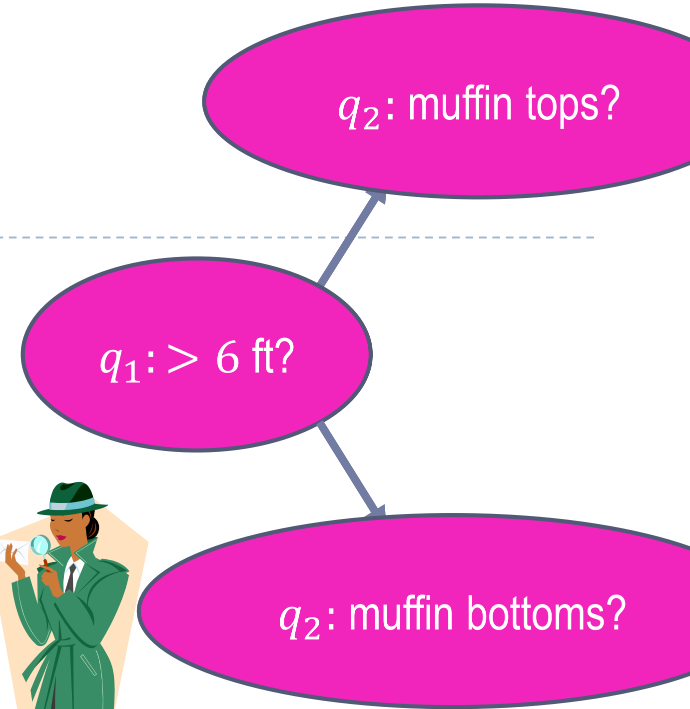
Cynthia Dwork, Microsoft Research

Adaptive Data Analysis



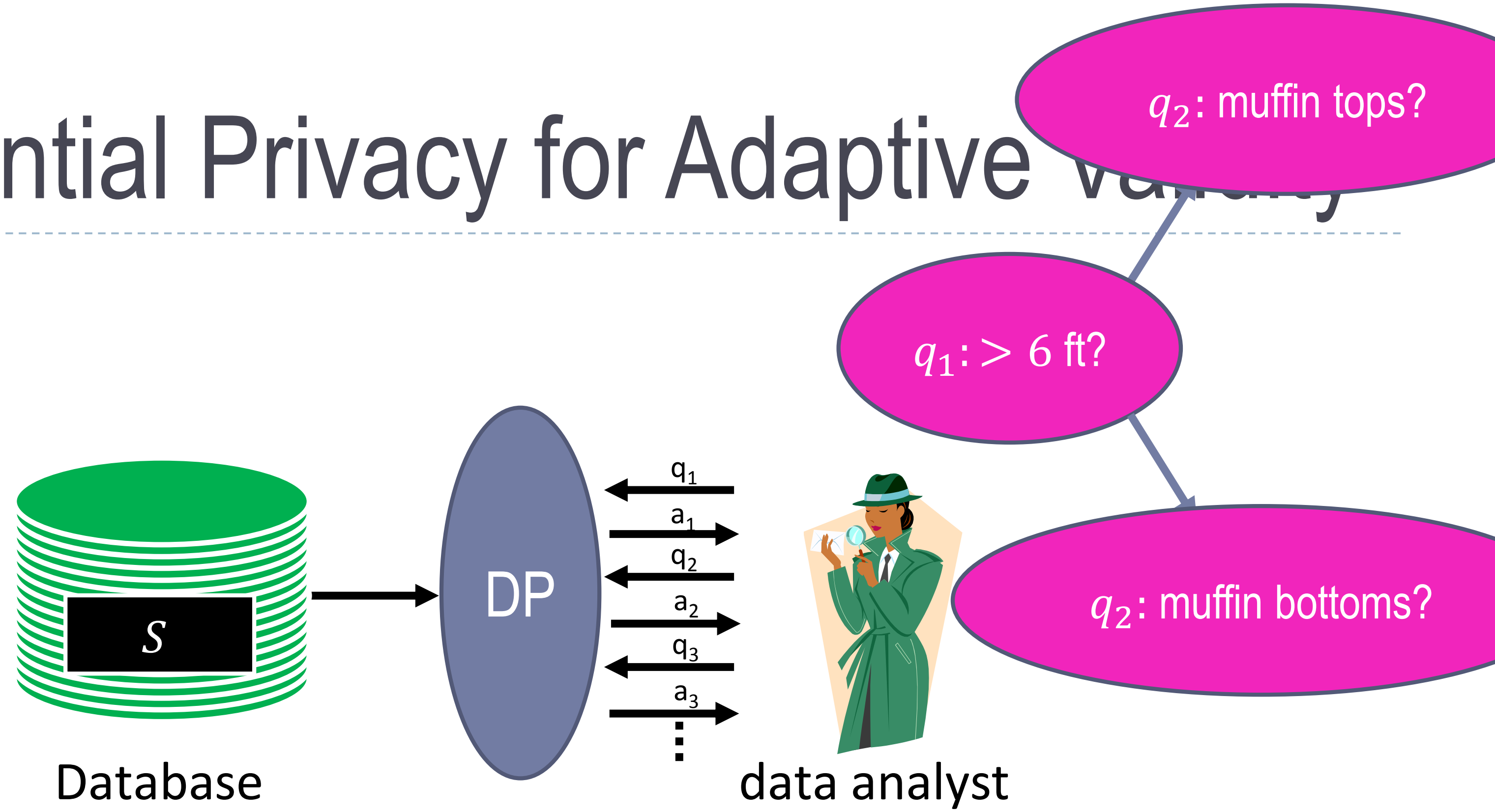
Database $S \sim D$

data analyst



- q_i depends on $a_1, a_2, \ldots, a_{i-1}$
- Worry: analyst finds a query for which the dataset is not representative of population; reports surprising discovery

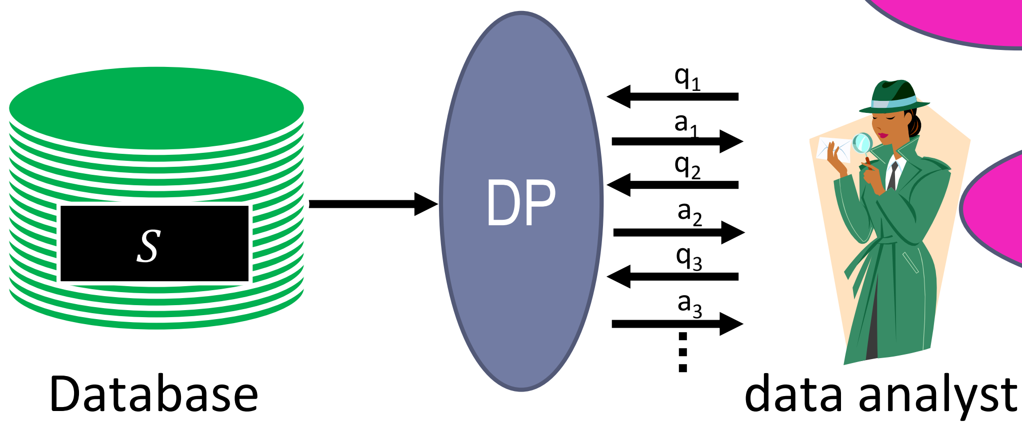
Differential Privacy for Adaptive Validity

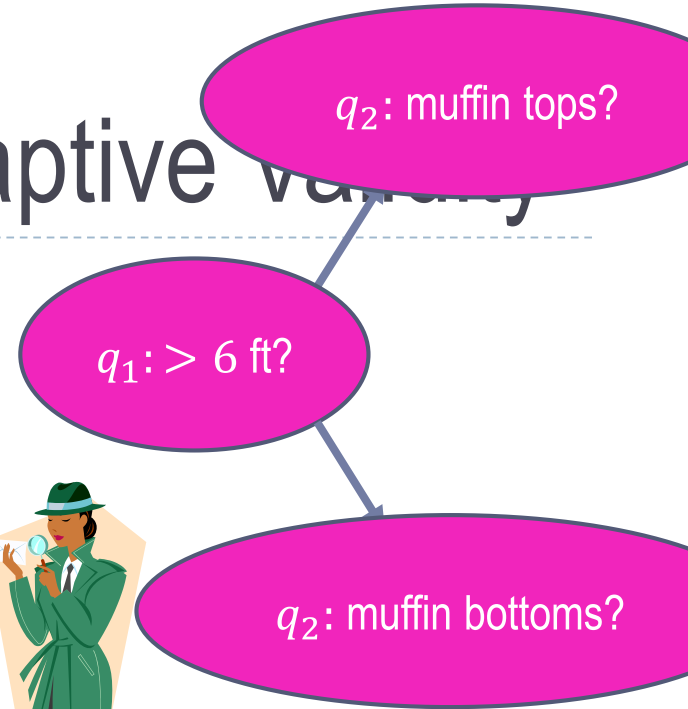


- q_i depends on $a_1, a_2, \dots, a_{i-1}$
- Differential privacy neutralizes risks incurred by adaptivity
  - Definition of privacy tailored to statistical analysis of large data sets

[D., Feldman, Hardt, Pitassi, Reingold, Roth '14]

Differential Privacy for Adaptive Validity





- q_i depends on $a_1, a_2, \ldots, a_{i-1}$
- Differential privacy neutralizes risks incurred by adaptivity
  - ∃ LARGE literature on DP algorithms for data analysis

[D., Feldman, Hardt, Pitassi, Reingold, Roth ’14]

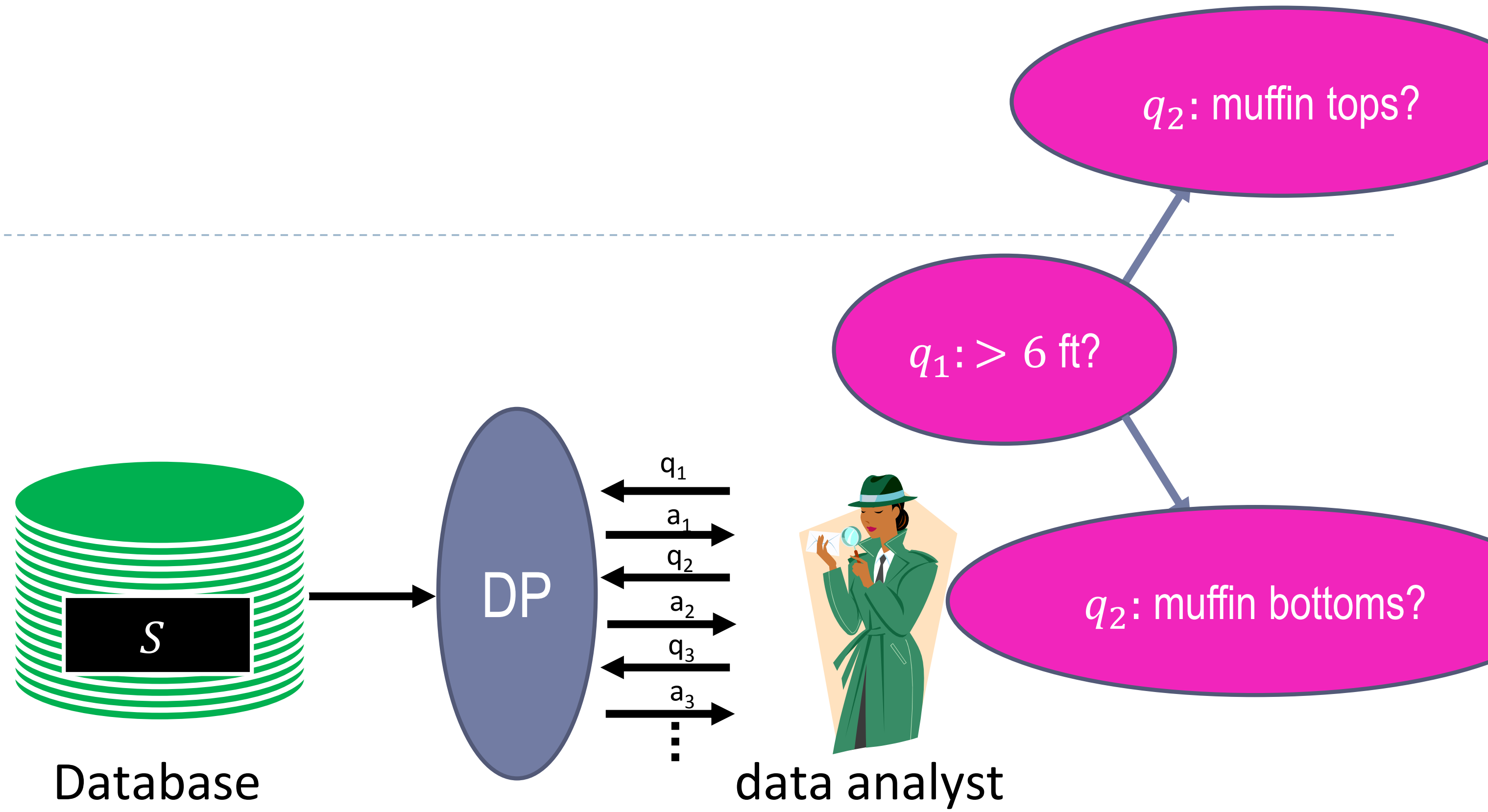
Some Intuition

- Fix a query, eg, “What fraction of population is over 6 feet tall?”
- Almost all large datasets will give an approximately correct reply
  - Most datasets are representative with respect to this query

- If, in the process of adaptive exploration, the analyst finds a query for which the dataset is not representative, then she must have “learned something significant” about the dataset.
  - Preserving the “privacy” of the data may prevent over-fitting.

Intuition After Nati's Talk

- Differential Privacy: The outcome of any analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset.
  - This is a stability requirement.
  - Gave rise to the folklore that differential privacy yields generalizability.
  - But we will be able to say something stronger.



- q_i depends on $a_1, a_2, \ldots, a_{i-1}$
- Differential privacy neutralizes risks incurred by adaptivity
  - E.g., for statistical queries: whp $|E_S[A(S)] - E_P[A(S)]| < \tau$
  - High probability is important for handling many queries

[D., Feldman, Hardt, Pitassi, Reingold, Roth '14]

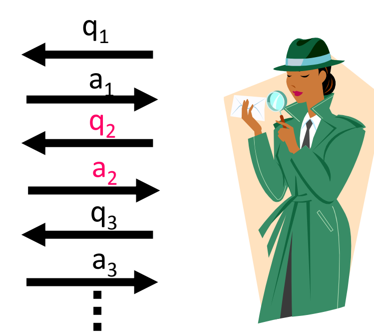
Formalization

- Data sets $S \in X^n$; $S \sim D$
- Queries $q: X^n \to Y$
- Algorithms that choose queries and output results
  - $A_1 = q_1$ (trivial choice), outputs $(q_1, q_1(S))$
  - A_i: $X^n \times Y_1 \times \cdots \times Y_{i-1} \to Y_i$ where
    - $q_i = C_i(y_1, \ldots, y_{i-1})$
    - $A_i(S, y_1, \ldots, y_{i-1}) = \big(q_i, q_i(S)\big) = (q_i, a_i)$
- $H \stackrel{\text{def}}{=} \{(S,q) \mid q(S) \text{ not representative wrt } D\}$
  - $\forall (y_1, \ldots, y_{i-1}) \Pr_S [(S, q_i) \in H] \leq \beta_i$
- We want: $\Pr[\big(\boldsymbol{S}, C_i(\boldsymbol{S})\big) \in H]$ to be similar
  - $q_i(S)$ should generalize even when q_i chosen as a function of S

Choose new query based on history of observations

Output chosen query and its response on S

$q_i(S)$ fails to generalize



Differential Privacy [D.,McSherry,Nissim,Smith ‘06]

M gives ϵ-differential privacy if for all pairs of adjacent data sets S, S', and all events T

$$\Pr[M(S) \in T] \leq e^{\epsilon} \Pr[M(S') \in T]$$

Randomness introduced by M

Differential Privacy [D.,McSherry,Nissim,Smith '06]

M gives ϵ-differential privacy if for all pairs of adjacent data sets S, S', and all events T

$$\Pr[M(S) \in T] \leq e^{\epsilon} \Pr[M(S') \in T]$$

For random variables $\boldsymbol{X}, \boldsymbol{Y}$ over X, the max-divergence of $\boldsymbol{X}$ from $\boldsymbol{Y}$ is given by

$$D_\infty(\boldsymbol{X}||\boldsymbol{Y}) = \log \max_{x \in X} \frac{\Pr[\boldsymbol{X} = x]}{\Pr[\boldsymbol{Y} = x]}$$

Then ϵ-DP equivalent to $D_\infty(M(S)||M(S')) \leq \epsilon$.

Differential Privacy [D.,McSherry,Nissim,Smith '06]

M gives ϵ-differential privacy if for all pairs of adjacent data sets S, S', and all events T

$$\Pr[M(S) \in T] \leq e^{\epsilon} \Pr[M(S') \in T]$$

For random variables $\boldsymbol{X}, \boldsymbol{Y}$ over X, the max-divergence of $\boldsymbol{X}$ from $\boldsymbol{Y}$ is given by

$$D_\infty(\boldsymbol{X}||\boldsymbol{Y}) = \log \max_{x \in X} \frac{\Pr[\boldsymbol{X} = x]}{\Pr[\boldsymbol{Y} = x]}$$

Then ϵ-DP equivalent to $D_\infty(M(S)||M(S')) \leq \epsilon$.

Closed under post-processing: $D_\infty(A(M(S))||A(M(S'))) \leq \epsilon$.

Differential Privacy [D.,McSherry,Nissim,Smith '06]

M gives ϵ-differential privacy if for all pairs of adjacent data sets S, S', and all events T

$$\Pr[M(S) \in T] \leq e^{\epsilon} \Pr[M(S') \in T]$$

For random variables $\boldsymbol{X}, \boldsymbol{Y}$ over X, the max-divergence of $\boldsymbol{X}$ from $\boldsymbol{Y}$ is given by

$$D_{\infty}(\boldsymbol{X}||\boldsymbol{Y}) = \log \max_{\mathrm{x} \in X} \frac{\Pr[\boldsymbol{X} = x]}{\Pr[\boldsymbol{Y} = x]}$$

Then ϵ-DP equivalent to $D_{\infty}(M(S)||M(S')) \leq \epsilon$.

Group Privacy: $\forall S, S''\ D_{\infty}\big(M(S)||M(S')\big) \leq \Delta(S, S'')\epsilon$.

Properties

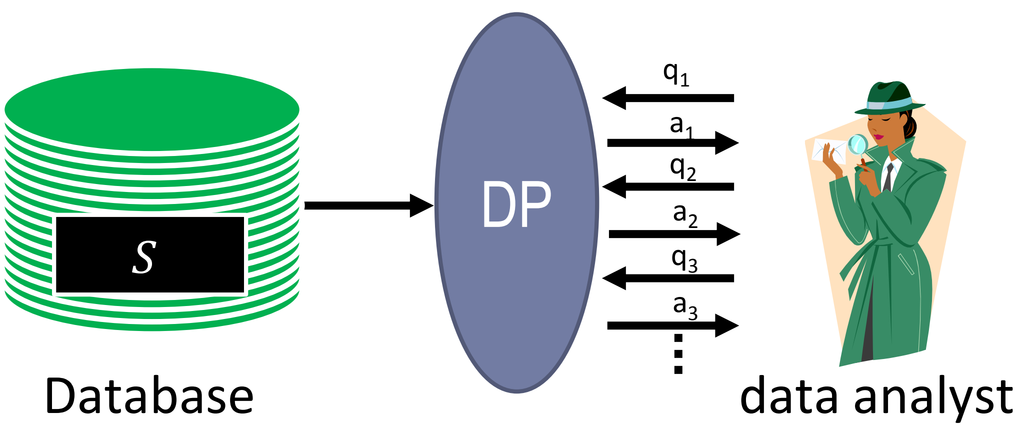
- Closed under post-processing
  - Max-divergence remains bounded
- Automatically yields group privacy
  - $k\epsilon$ for groups of size k
- Understand behavior under adaptive composition
  - Can bound cumulative privacy loss over multiple analyses
    - "The epsilons add up"
- Programmable
  - Complicated private analyses from simple private building blocks

The Power of Composition

- Lemma: The choice of q_i is differentially private.
  - Closure under post-processing.
- Inductive step (key): If q is chosen in a differentially private fashion with respect to S, then

$$\Pr[(\boldsymbol{S}, C(\boldsymbol{S})) \in H] \text{ is small}$$

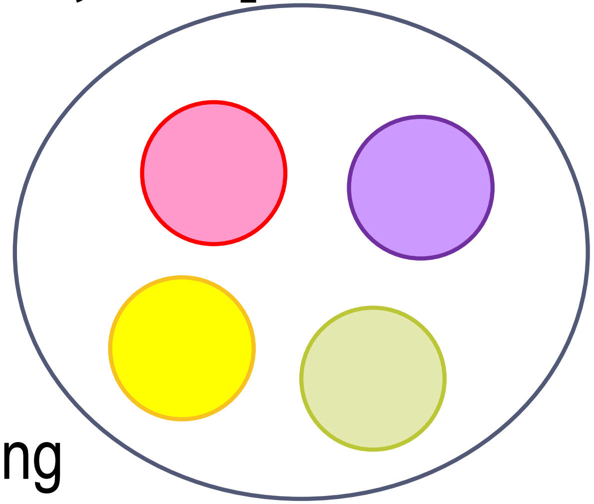
  - Sufficiency: union bound.



Description Length

- Let $A: X^n \to Y$.
- Description length of A is the cardinality of its range

If $\forall y \ \Pr_S\left[(S, y) \in H\right] \le \beta$, then $\Pr_S\left[\left(S, A(S)\right) \in H\right] \le |Y| \cdot \beta$



- Description length composes too.
  - Product: $\beta \cdot \Pi_i |Y_i|$
- And, morally, it is closed under post-processing
  - Once you fix the randomness of the post-processing algorithm

[D., Feldman, Hardt, Pitassi, Reingold, Roth '15]

Approximate max-divergence

β-approximate max-divergence of $\boldsymbol{X}$ from $\boldsymbol{Y}$

$$D_\infty^\beta(\boldsymbol{X}||\boldsymbol{Y}) = \log \max_{T\in X,\ \Pr[\boldsymbol{X}\in T]>\beta} \frac{\Pr[\boldsymbol{X} \in T] - \beta}{\Pr[\boldsymbol{Y} \in T]}$$

Approximate max-divergence

β-approximate max-divergence of $\boldsymbol{X}$ from $\boldsymbol{Y}$

$$D_\infty^\beta(\boldsymbol{X}||\boldsymbol{Y}) = \log \max_{T \in X,\ \Pr[\boldsymbol{X} \in T] > \beta} \frac{\Pr[\boldsymbol{X} \in T] - \beta}{\Pr[\boldsymbol{Y} \in T]}$$

We are interested in (with β, but too messy)

$$D_\infty((\boldsymbol{S}, A(\boldsymbol{S}))||\boldsymbol{S} \times A(\boldsymbol{S})) = \log \max_T \frac{\Pr[(\boldsymbol{S}, A(\boldsymbol{S})) \in T]}{\Pr[\boldsymbol{S} \times A(\boldsymbol{S}) \in T]}$$

Approximate max-divergence

β-approximate max-divergence of $\boldsymbol{X}$ from $\boldsymbol{Y}$

$$D_{\infty}^{\beta}(\boldsymbol{X}||\boldsymbol{Y}) = \log \max_{T \in X,\ \Pr[\boldsymbol{X} \in T] > \beta} \frac{\Pr[\boldsymbol{X} \in T] - \beta}{\Pr[\boldsymbol{Y} \in T]}$$

We are interested in (with β, but too messy)

$$D_{\infty}((\boldsymbol{S}, A(\boldsymbol{S}))||\boldsymbol{S} \times A(\boldsymbol{S})) = \log \max_{T} \frac{\Pr[(\boldsymbol{S}, A(\boldsymbol{S})) \in T]}{\Pr[\boldsymbol{S} \times A(\boldsymbol{S}) \in T]}$$

How much more likely is $A(S)$ to relate to S than to a fresh S'?

Captures the maximum amount of information that an output of an algorithm might reveal about its input

Unifying Concept: Max-Information

- $I_\infty^\beta(\boldsymbol{X};\boldsymbol{Y}) = D_\infty^\beta((\boldsymbol{X},\boldsymbol{Y})||\boldsymbol{X}\times\boldsymbol{Y})$
- We are interested in $I_\infty^\beta(\boldsymbol{S};A(\boldsymbol{S}))$
- Theorem: If $I_\infty^\beta\big(\boldsymbol{S};A(\boldsymbol{S})\big) \leq k$ then for any $T \subseteq X^n \times Y$
  - $\Pr\big[\big(\boldsymbol{S},A(\boldsymbol{S})\big) \in T\big] \leq 2^k \Pr[\boldsymbol{S}\times A(\boldsymbol{S}) \in T] + \beta$
  - So $\Pr\big[\big(\boldsymbol{S},A(\boldsymbol{S})\big) \in H\big] \leq 2^k \max_{y\in Y} \Pr[(\boldsymbol{S},y) \in H] + \beta$!

[D., Feldman, Hardt, Pitassi, Reingold, Roth '15]

Unifying Concept: Max-Information

- $I_\infty^\beta(\boldsymbol{X};\boldsymbol{Y}) = D_\infty^\beta((\boldsymbol{X},\boldsymbol{Y})||\boldsymbol{X}\times\boldsymbol{Y})$
- We are interested in $I_\infty^\beta(\boldsymbol{S};A(\boldsymbol{S}))$
- Theorem: If $I_\infty^\beta\big(\boldsymbol{S};A(\boldsymbol{S})\big) \le k$ then for any $T \subseteq X^n \times Y$
  - $\Pr\big[\big(\boldsymbol{S},A(\boldsymbol{S})\big) \in T\big] \le 2^k \Pr[\boldsymbol{S}\times A(\boldsymbol{S}) \in T] + \beta$
  - So $\Pr\big[\big(\boldsymbol{S},A(\boldsymbol{S})\big) \in H\big] \le 2^k \max_{y\in Y} \Pr[(\boldsymbol{S},y) \in H] + \beta$!
- Max-Information composes and is closed under post-processing
- For ϵ-DP A: $I_\infty(A,n) \le \epsilon n \log_2 e$. Better bounds for $I_\infty^\beta(A,n)$.
- $I_\infty^\beta(A,n) \le \log\left(\frac{|Y|}{\beta}\right)$

Bound on worst case approximate max info for any distribution on n-element databases

[D., Feldman, Hardt, Pitassi, Reingold, Roth '15]

Abstract is Good

- Focusing on *properties* is powerful
  - Completely universal approach to validity of adaptive analysis
    - DP, small description length, low max-information
  - Large numbers of arbitrary adaptively chosen computations
    - Closure under post-processing and composition

Long Live the Dataset!

- Leaking information slowly prolongs the lifetime of the system

- Similar to the situation with privacy for the sake of privacy
  - To avoid too much cumulative loss, answer with smaller values of ϵ
  - Essential: Fundamental Law of Information Leakage
    - Overly accurate estimates of too many statistics is blatantly non-private.
    - Dealer's choice

- Conjecture: The same is true for adaptivity.
  - Failure to control cumulative max-info leads to failure to generalize
  - Important policy Implications!
  - Supporting evidence: Hardt-Ullman queries

Thank you!

NIPS Workshop on Adaptive Data Analysis, Montreal, 12/11/15