Post-processing outputs for better utility

CompSci 590.03

Instructor: Ashwin Machanavajjhala

Duke UNIVERSITY

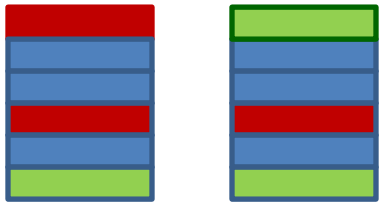
Announcement

- Project proposal submission deadline is **Fri, Oct 12 noon**.

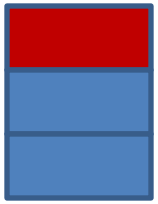
Recap: Differential Privacy

For every pair of inputs that differ in one value

For every output …



D_1 D_2

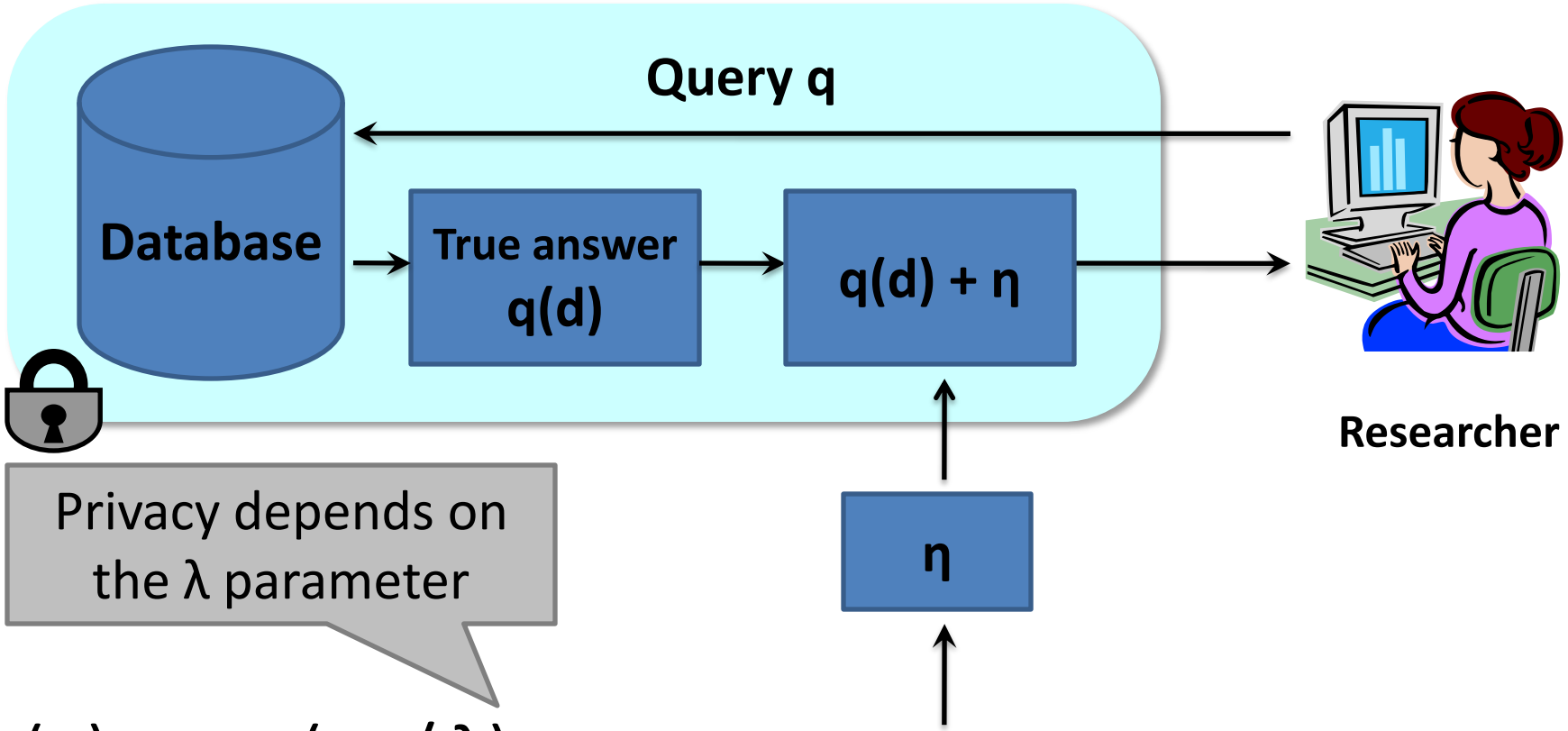


O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

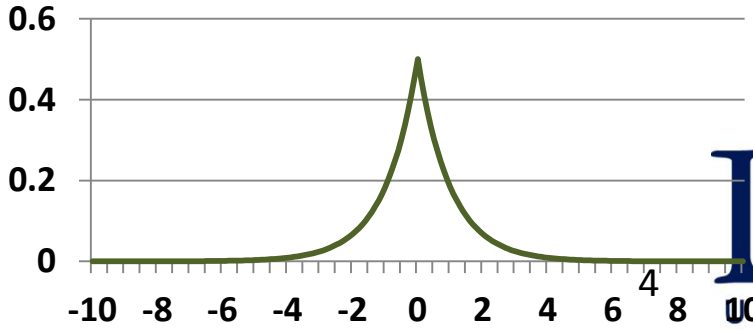
$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \qquad (\varepsilon>0)$$

Recap: Laplacian Distribution



$h(\eta) \alpha \exp(-\eta / \lambda)$

Laplace Distribution – Lap(λ)



Mean: 0,
Variance: $2\lambda^2$

Duke UNIVERSITY

Recap: Laplace Mechanism

Thm: If **sensitivity** of the query is **S**, then the following guarantees ε-differential privacy.

$$\lambda = S/\varepsilon$$

Sensitivity: Smallest number s.t. for any d, d' differing in one entry,

$$|| q(d) - q(d') || \leq S(q)$$

Histogram query: Sensitivity = 2

- Variance / error on each entry = $2x4/\varepsilon^2 =$ **$O(1/\varepsilon^2)$**

This class

- What is the optimal method to answer a batch of queries?

How to answer a batch of queries?

- Database of values {x1, x2, …, xk}

- Query Set:
  - Value of x1 $\eta 1 = x1 + \delta 1$
  - Value of x2 $\eta 2 = x2 + \delta 2$
  - Value of x1 + x2 $\eta 3 = x1 + x2 + \delta 3$

- But we know that η1 and η2 should sum up to η3!

Two Approaches

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A

  - **Universal Histograms**
  - **Wavelet Mechanism**
  - **Matrix Mechanism**

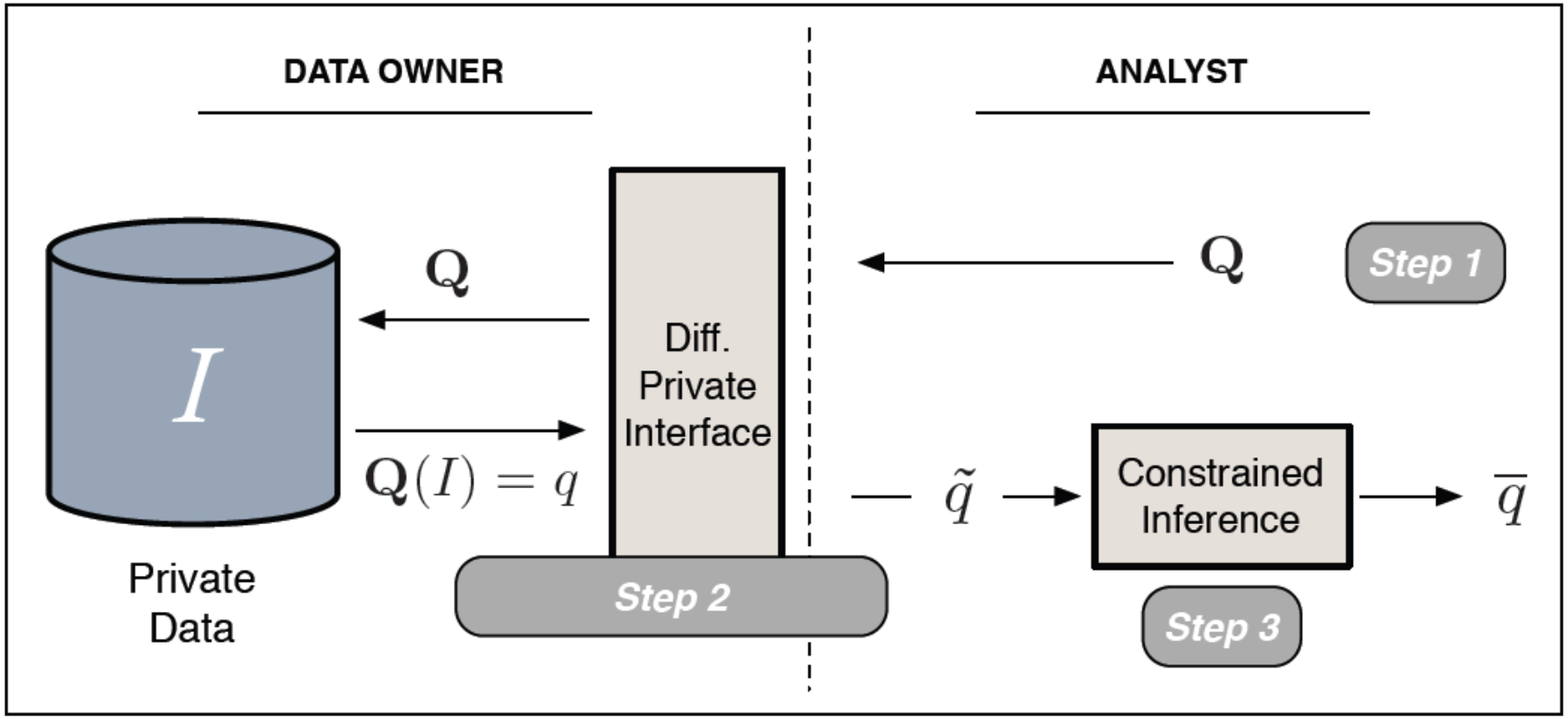
Two Approaches

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A

  - **Universal Histograms**
  - **Wavelet Mechanism**
  - **Matrix Mechanism**

Constrained Inference



Duke UNIVERSITY

Constrained Inference

- Let x1 and x2 be the original values. We observe noisy values η1, η2 and η3
- We would like to reconstruct the best estimators y1 (for x1/) and y2 (for x2) from the noisy values.

- That is, we want to find the values of y1, y2 such that:

$$\min (y1-\eta1)^2 + (y2 - \eta2)^2 + (y3 - \eta3)^2$$
$$\text{s.t., } y1 + y2 = y3$$

Constrained Inference [Hay et al VLDB 10]

DEFINITION 2.4 (MINIMUM L_2 SOLUTION). *Let* $\mathbf{Q}$ *be a query sequence with constraints* $\gamma_{\mathbf{Q}}$. *Given a noisy query sequence* $\tilde{q} = \tilde{\mathbf{Q}}(I)$, *a minimum* L_2 *solution, denoted* $\overline{q}$, *is a vector* $\overline{q}$ *that satisfies the constraints* $\gamma_{\mathbf{Q}}$ *and at the same time minimizes* $||\tilde{q} - \overline{q}||_2$.

Sorted Unattributed Histograms

- Counts of diseases
  - (without associating a particular count to the corresponding disease)
- Degree sequence: List of node degrees
  - (without associating a degree to a particular node)

- Constraint: The values are sorted

Sorted Unattributed Histograms

True Values 20, 10, 8, 8, 8, 5, 3, 2

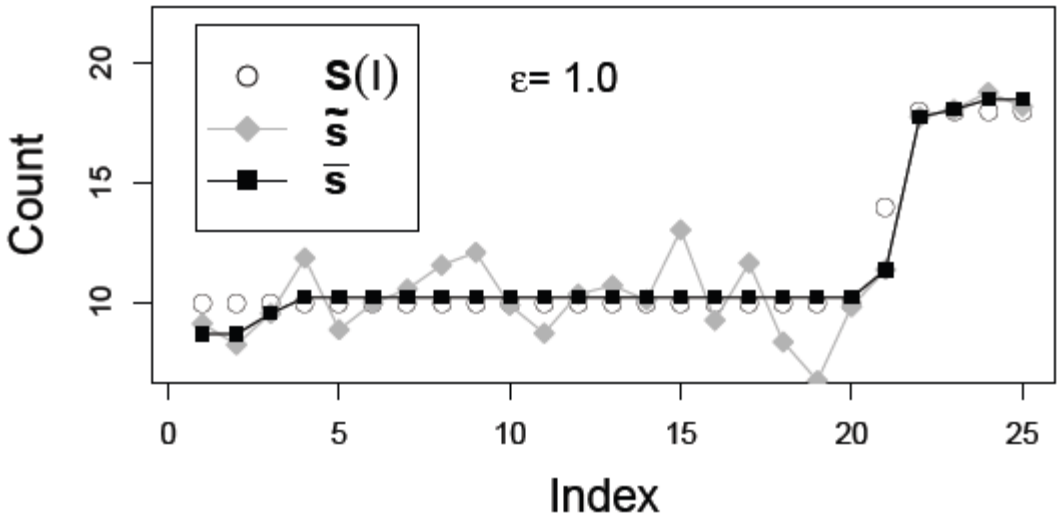
Noisy Values 25, 9, 13, 7, 10, 6, 3, 1 (noise from Lap(1/ε))

$$minimize \left|\left|\tilde{s} - \bar{s}\right|\right|_2 \quad s.t. \forall i, \bar{s}_i \leq \bar{s}_{i-1}$$

$$\bar{s}_k = \min_{j \in [k,n]} \max_{i \in [1,j]} \frac{\left(\tilde{s}_i + \tilde{s}_{i+1} + \cdots + \tilde{s}_j\right)}{j - i + 1}$$

Proof:?

Sorted Unattributed Histograms



Duke UNIVERSITY

Sorted Unattributed Histograms

- n: number of values in the histogram
- d: number of distinct values in the histogram
- n_i: number of times i^{th} distinct value appears in the histogram.

THEOREM 2. *There exist constants c_1 and c_2 independent of n and d such that*

$$error(\overline{\mathbf{S}}) \leq \sum_{i=1}^{d} \frac{c_1 \log^3 n_i + c_2}{\epsilon^2}$$

Thus $error(\overline{\mathbf{S}}) = O(d \log^3 n/\epsilon^2)$ whereas $error(\tilde{\mathbf{S}}) = \Theta(n/\epsilon^2)$.

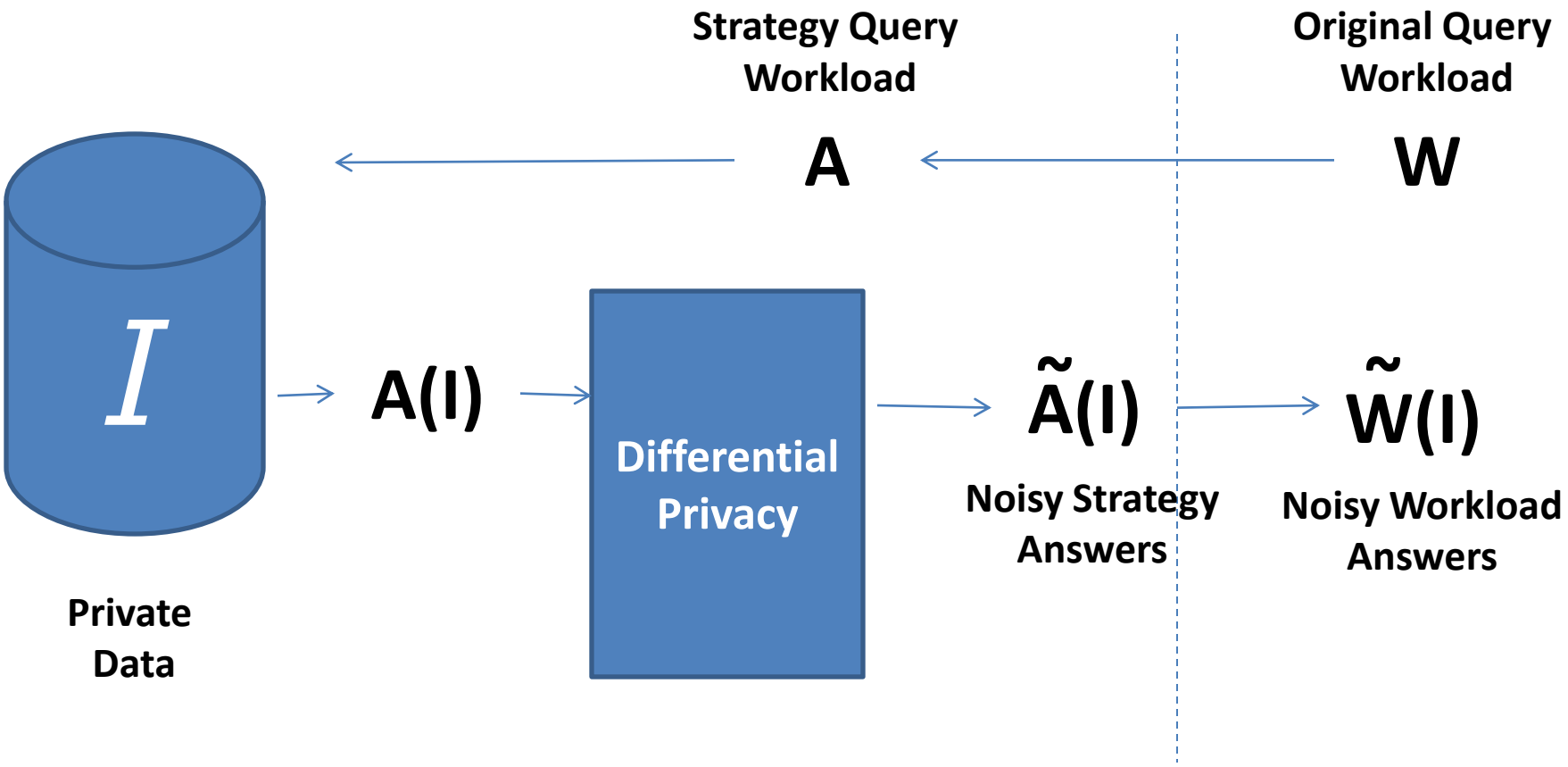
Two Approaches

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A

  - **Universal Histograms**
  - **Wavelet Mechanism**
  - **Matrix Mechanism**

Query Strategy



Duke UNIVERSITY

Range Queries

- Given a set of values {x1, x2, …, xn}
- Range query: q(j,k) = xj + … + xk

Q: Suppose we want to answer all range queries?

Strategy 1: Answer all range queries using Laplace mechanism

- $O(n^2/\varepsilon^2)$ total error.
- May reduce using constrained optimization …

Range Queries

- Given a set of values {x1, x2, …, xn}
- Range query: q(j,k) = xj + … + xk

Q: Suppose we want to answer all range queries?

Strategy 1: Answer all range queries using Laplace mechanism

- Sensitivity = $O(n^2)$
- $O(n^4/\varepsilon^2)$ total error across all range queries.
- May reduce using constrained optimization …

Range Queries

- Given a set of values {x1, x2, …, xn}
- Range query: q(j,k) = xj + … + xk

Q: Suppose we want to answer all range queries?

Strategy 2: Answer all xi queries using Laplace mechanism
Answer range queries using noisy xi values.

- $O(1/\varepsilon^2)$ error for each xi.
- Error(q(1,n)) = $O(n/\varepsilon^2)$
- Total error on all range queries : $O(n^3/\varepsilon^2)$

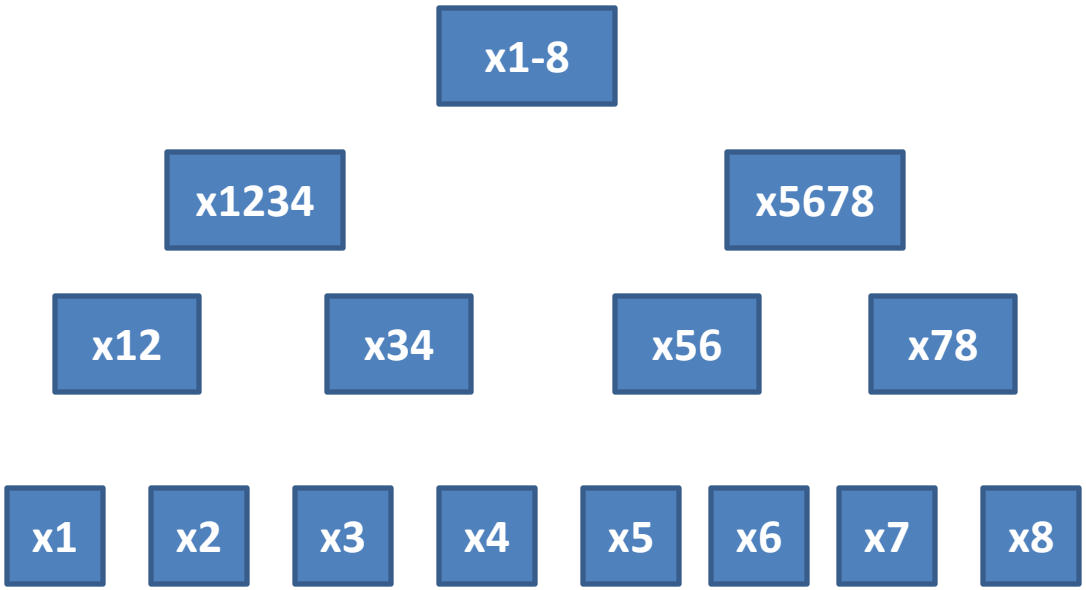
Universal Histograms for Range Queries

[Hay et al VLDB 2010]

Strategy 3:

Answer *sufficient statistics* using Laplace mechanism

Answer range queries using noisy sufficient statistics.



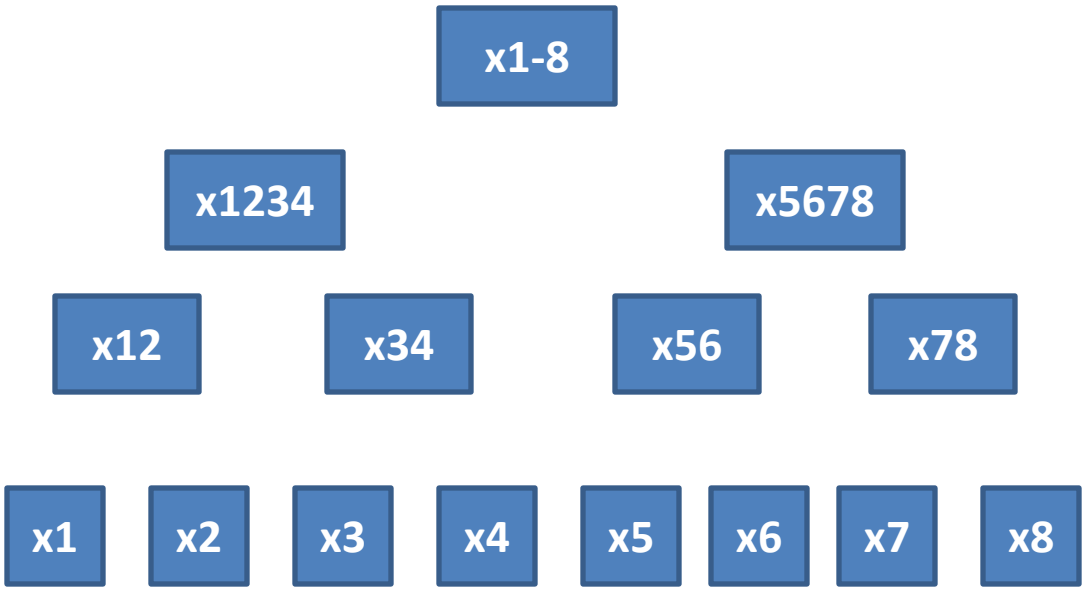
Duke UNIVERSITY

Universal Histograms for Range Queries

- Sensitivity: log n
- q(2,6) = x2+x3+x4+x5+x6
  = x2 + x34 + x56

Error = $2 \times 5\log^2 n/\varepsilon^2$

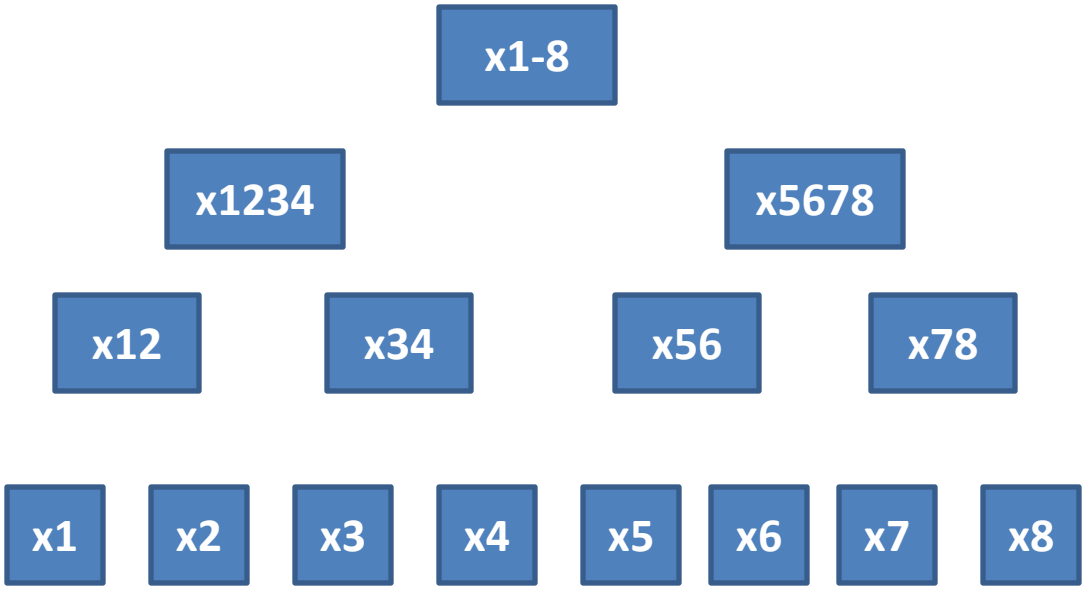
Error = $2 \times 3\log^2 n/\varepsilon^2$



Duke UNIVERSITY

Universal Histograms for Range Queries

- Every range query can be answered by summing at most log n different noisy answers
- Maximum error on any range query = $O(\log^3 n / \varepsilon^2)$
- Total error on all range queries = $O(n^2 \log^3 n / \varepsilon^2)$



Duke UNIVERSITY

Universal Histograms & Constrained Inference

[Hay et al VLDB 2010]

- Can further reduce the error by enforcing constraints
  x1234 = x12 + x34 = x1 + x2 + x3 + x4

$$minimize \sum_{v} \left(\tilde{c}(v) - \bar{c}(v)\right)^2$$

$$s.t.\ \bar{c}(v) = \sum_{u \in child(v)} \bar{c}(u)$$

- 2-pass algorithm to compute a consistent version of the counts

Universal Histograms & Constrained Inference

[Hay et al VLDB 2010]

- Pass 1: (Bottom Up)

$$z(v) = \begin{cases} \tilde{c}(v), & if\ leaf\ node \\ \alpha \cdot c(v) + (1-\alpha) \sum_{u=child(v)} z(u) \end{cases}$$

- Pass 2: (Top down)

$$\bar{c}(v) = \begin{cases} z(v), & if\ root\ node \\ z(v) + \frac{1}{2}\left(\bar{c}(v) - \sum_{u=child(v)} z(u) \right) \end{cases}$$

Universal Histograms & Constrained Inference

- Resulting consistent counts
  - Have lower error than noisy counts (upto 10 times smaller in some cases)
  - Unbiased estimators
  - Have the least error amongst all unbiased estimators

Next Class

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A

  - **Universal Histograms**
  - **Wavelet Mechanism**
  - **Matrix Mechanism**