Building complex DP algorithms using composition

Privacy & Fairness in Data Science

CompSci 590.01 Fall 2018

DUKE COMPUTER SCIENCE

Outline

- Recap
  - Laplace Mechanism
- Composition Theorems
- Optimizing accuracy of DP algorithms
  - Utilizing Parallel Composition
  - Postprocessing & Inference
  - Strategy Selection
  - Data dependent noise

Differential Privacy

[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …

D_1 D_2 O

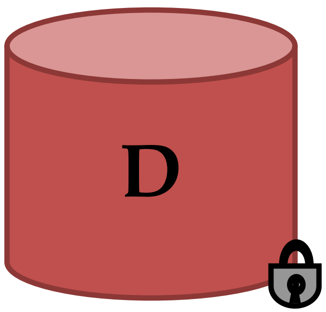
Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\forall \Omega \in \text{range}(\mathrm{A}),\ \ln\left(\frac{\Pr[A(D_1) \in \Omega]}{\Pr[A(D_2) \in \Omega]}\right) \leq \varepsilon, \qquad \varepsilon > 0$$

Laplace mechanism

e.g., COUNT

Aggregate Query: q

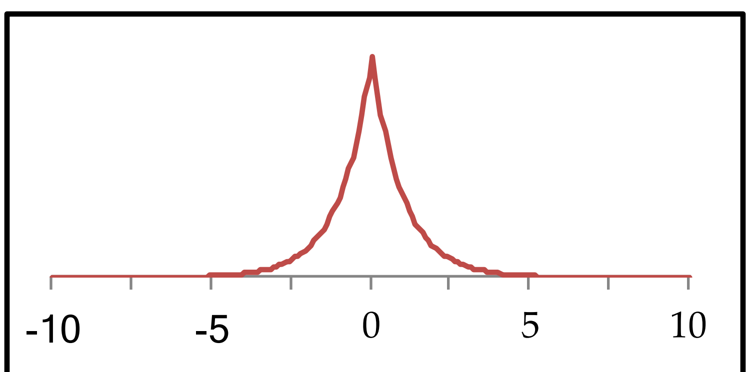


Noisy Answer

Analyst

Private Database

$$\tilde{q}(D) = q(D) + \mathrm{Lap}\left(\frac{S(q)}{\varepsilon}\right)$$



Sensitivity

Laplace Mechanism

Theorems:

$$E\left(\left(\tilde{q}(D) - q(D)\right)^2\right) = 2\left(\frac{S(q)}{\varepsilon}\right)^2$$

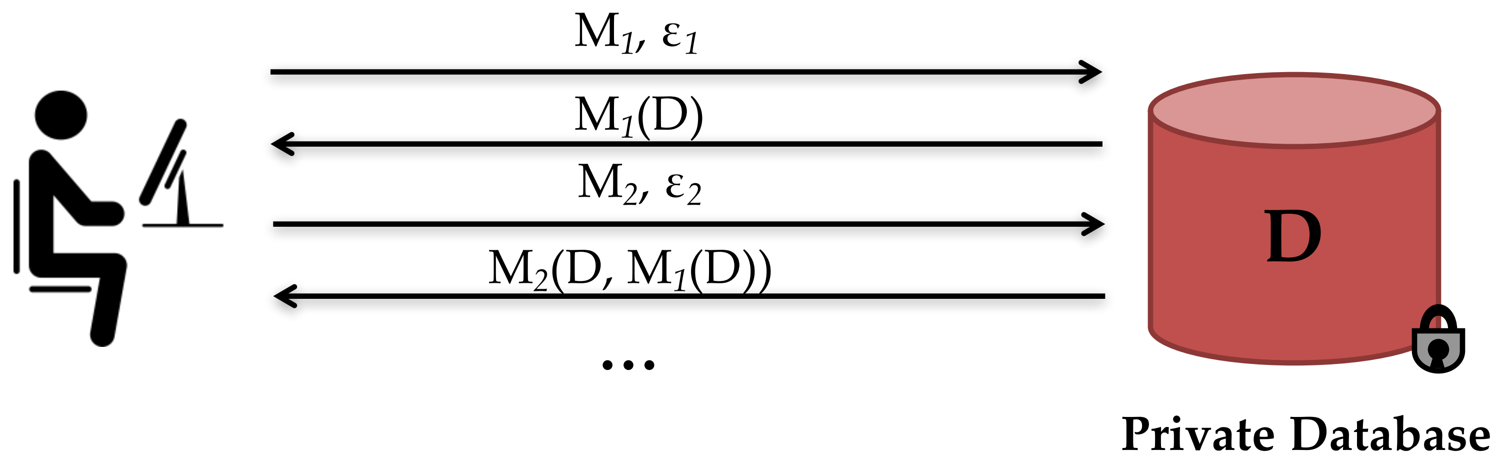
Error is *data independent*
Depends on q and ε, but not on D

$$Pr\left[|\tilde{q}(D) - q(D)| \geq \frac{S(q)}{\varepsilon}\ln\left(\frac{1}{\delta}\right)\right] \leq \delta$$

Outline

- Recap
  - Laplace Mechanism
- Composition Theorems
- Optimizing accuracy of DP algorithms
  - Utilizing Parallel Composition
  - Postprocessing & Inference
  - Strategy Selection
  - Data dependent noise

Sequential Composition

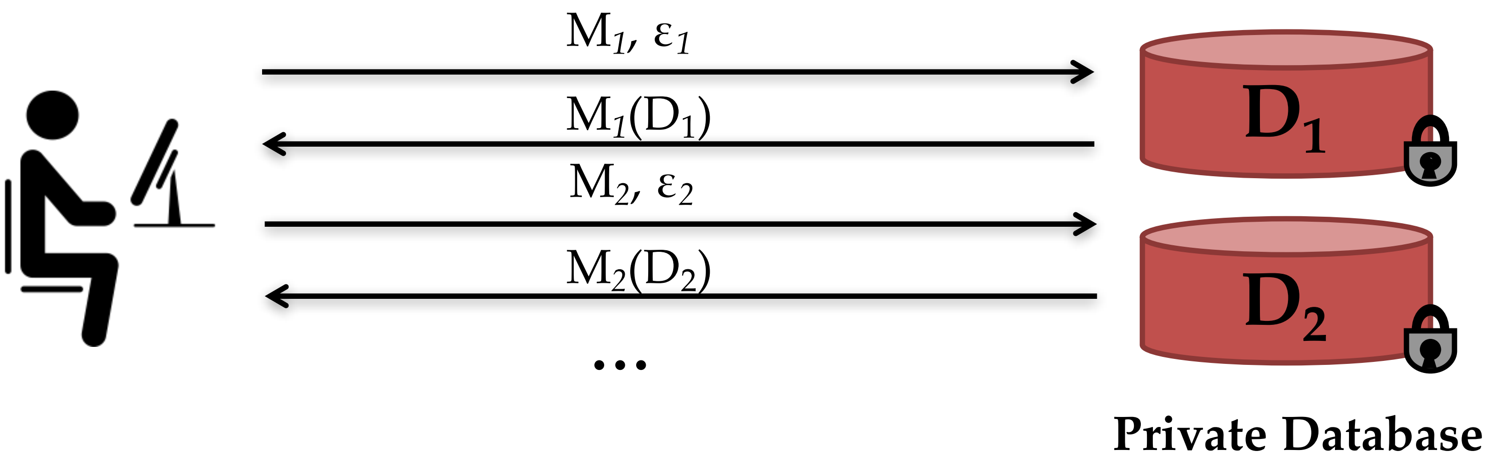


- If $M_1, M_2, \ldots, M_k$ are algorithms that access a private database D such that each M_i satisfies ε_i -differential privacy,

  then the combination of their outputs satisfies ε-differential privacy with

$$\varepsilon = \varepsilon_1 + \ldots + \varepsilon_k$$

Parallel Composition

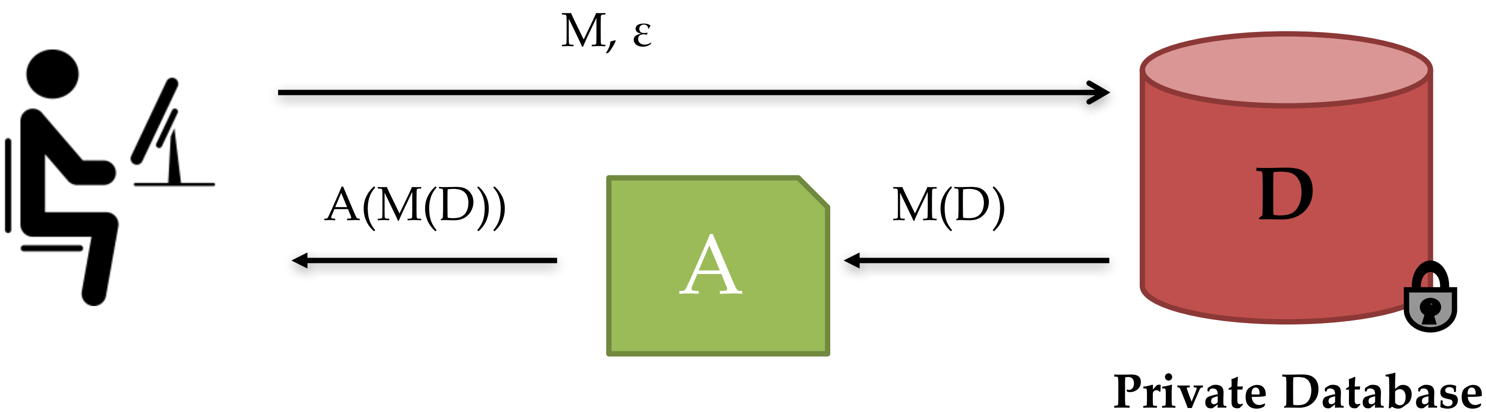


- If $M_1, M_2, \ldots, M_k$ are algorithms that access are algorithms that access disjoint databases $D_1, D_2, \ldots, D_k$ such that each M_i satisfies ε_i -differential privacy,

  then the combination of their outputs satisfies ε-differential privacy with

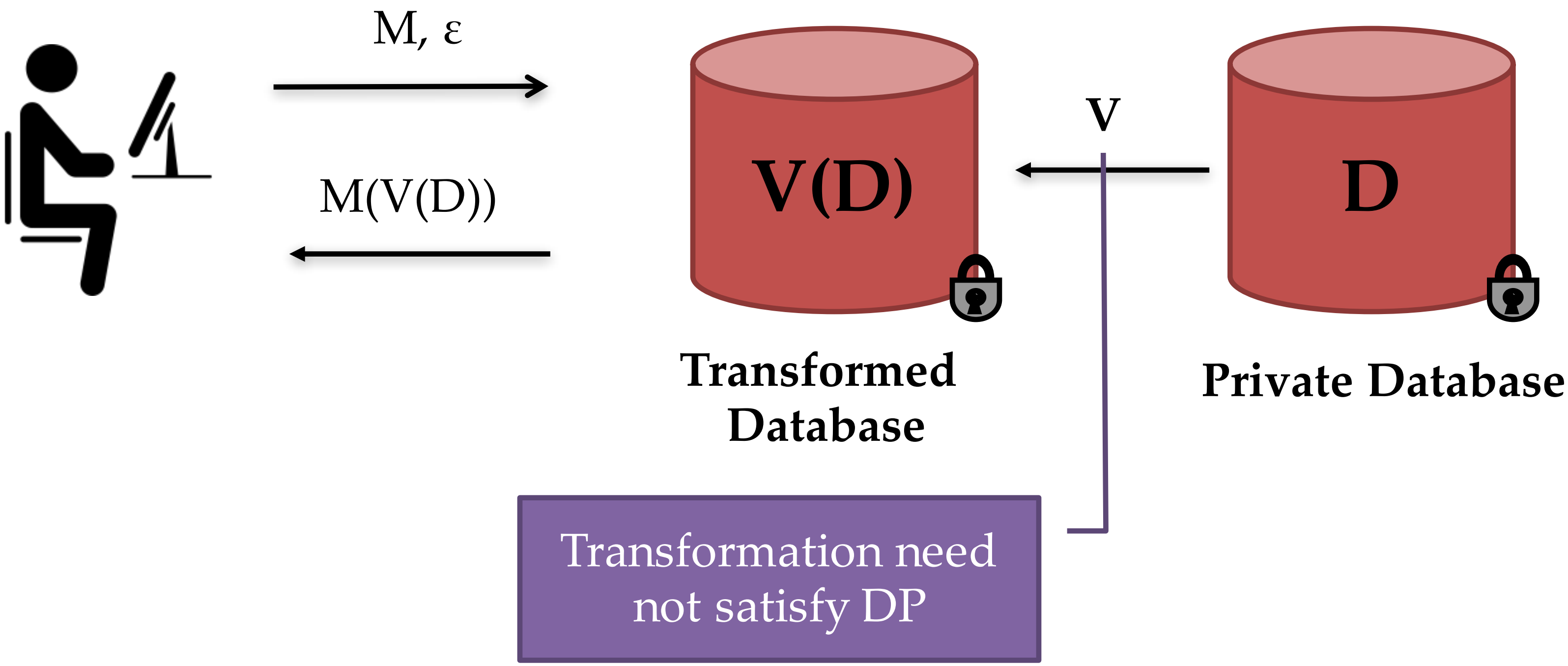
$$\varepsilon = \max(\varepsilon_1, \ldots, \varepsilon_k)$$

Postprocessing



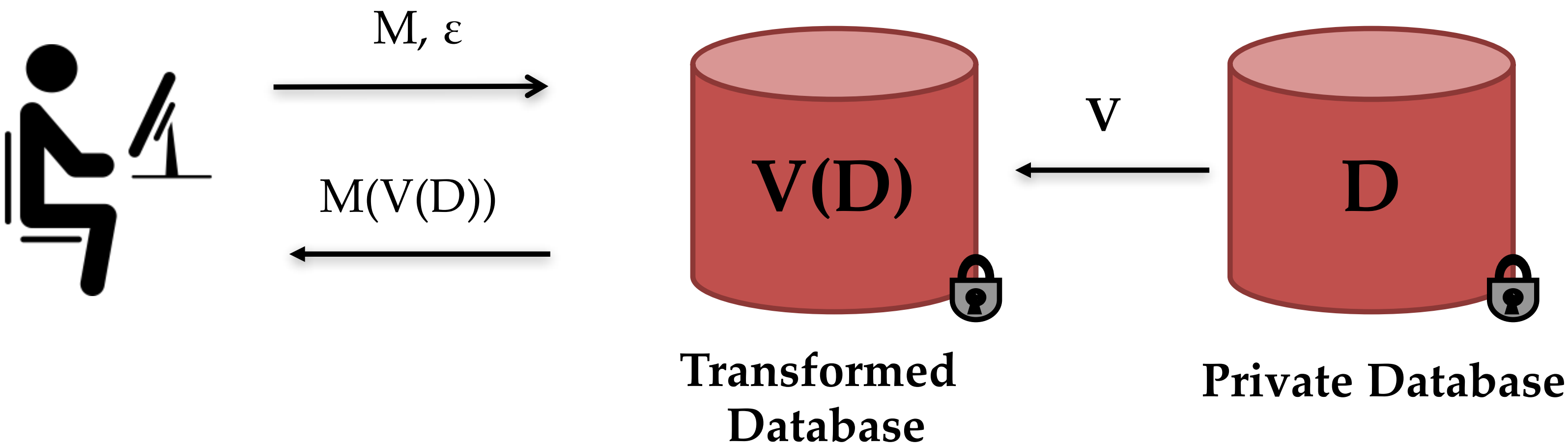
- If M is an ε-differentially private algorithm, any additional post-processing $A \circ M$ also satisfies ε-differential privacy.

Transformations & Stability



- σ_V: Stability of the transformation
  - Maximum number of rows in V that can change due to changing a single row in D

Transformations & Stability



- Executing an ε-differentially private algorithm M on a transformation of a database V(D) satisfies $\varepsilon \cdot \sigma_V$-differential privacy.

- σ_V: Stability of the transformation
  – Maximum number of rows in V that can change due to changing a single row in D

Transformations & Stability

- V_1: For each row (x1, x2, x3) → (x1, x2+x3)

  Stability = 1

- V_2: Each row in D is a tweet (id, {words}). For each row in D, generate k rows with first k words {(id, $word_1$), …, (id, $word_k$)}

  Stability = k

- V_3: Sample each row with probability p.

  Stability = 1 … but can prove $2p\varepsilon$ -differential privacy*

*Adam Smith, [Differential Privacy and Secrecy of the Sample]

Outline

- Recap
  - Laplace Mechanism

- Composition Theorems

- Optimizing accuracy of DP algorithms
  - Utilizing Parallel Composition
  - Postprocessing & Inference
  - Strategy Selection
  - Data dependent noise

Problem

Sex	Height	Weight
M	6′2″	210
F	5′3″	190
F	5′9″	160
M	5′3″	180
M	6′7″	250

Queries:

- # Males with BMI < 25
- # Males
- # Females with BMI < 25
- # Females

- Design an ε-differentially private algorithm that can answer all these questions.
- What is the total error?

Algorithm 1

Return:

- # Males with BMI < 25 + Lap($4/\varepsilon$)
- # Males + Lap($4/\varepsilon$)
- # Females with BMI < 25 + Lap($4/\varepsilon$)
- # Females + Lap($4/\varepsilon$)

Privacy

- BMI can be computed by transforming each row (s, h, w) → (s, bmi). This is stability 1.
- Sensitivity of count = 1. So each query is answered using a ε/4-DP algorithm.
- By sequential composition, we get ε-DP.

Utility

Error:

$$\sum E\left(\left(\tilde{q}(D) - q(D)\right)^2\right)$$

Total Error:

$$2\left(\frac{4}{\varepsilon}\right)^2 \times 4 = \frac{128}{\varepsilon^2}$$

Algorithm 2

Compute:

- $\widetilde{q_1}$ = # Males with BMI < 25 + Lap(1/ε)
- $\widetilde{q_2}$ = # Males with BMI > 25 + Lap(1/ε)
- $\widetilde{q_3}$ = # Females with BMI < 25 + Lap(1/ε)
- $\widetilde{q_4}$ = # Females with BMI > 25 + Lap(1/ε)

Return

- $\widetilde{q_1}$, $\widetilde{q_1}+\widetilde{q_2}$, $\widetilde{q_3}$, $\widetilde{q_3}+\widetilde{q_4}$

Privacy

- Sensitivity of count = 1. So each query is answered using a ε-DP algorithm.

- q_1, q_2, q_3, q_4 are counts on disjoint portions of the database. Thus by *parallel composition* releasing $\widetilde{q_1}, \widetilde{q_2}, \widetilde{q_3}, \widetilde{q_4}$ satisfies ε-DP.

- By the *postprocessing theorem*, releasing $\widetilde{q_1}, \widetilde{q_1}+\widetilde{q_2}, \widetilde{q_3}, \widetilde{q_3}+\widetilde{q_4}$ also satisfies ε-DP.

Utility

Error:

$$\sum E\left(\left(\tilde{q}(D) - q(D)\right)^2\right)$$

Total Error:

$$2\left(\frac{1}{\varepsilon}\right)^2 + 2\cdot 2\left(\frac{1}{\varepsilon}\right)^2 + 2\left(\frac{1}{\varepsilon}\right)^2 + 2\cdot 2\left(\frac{1}{\varepsilon}\right)^2 = \frac{12}{\varepsilon^2}$$

$\widetilde{q_1}$ $\quad$ $\widetilde{q_1} + \widetilde{q_2}$ $\quad$ $\widetilde{q_3}$ $\quad$ $\widetilde{q_3} + \widetilde{q_4}$

Utility

Tighter privacy analysis gives better accuracy for the same level of privacy

Total Error:

$$2\left(\frac{1}{\varepsilon}\right)^2 + 2\cdot 2\left(\frac{1}{\varepsilon}\right)^2 + 2\left(\frac{1}{\varepsilon}\right)^2 + 2\cdot 2\left(\frac{1}{\varepsilon}\right)^2 = \frac{12}{\varepsilon^2}$$

$\widetilde{q_1}$ $\widetilde{q_1} + \widetilde{q_2}$ $\widetilde{q_3}$ $\widetilde{q_3} + \widetilde{q_4}$

Generalized Sensitivity

- Let $f: \mathcal{D} \to \mathbb{R}^d$ be a function that outputs a vector of d real numbers. The sensitivity of f is given by:

$$S(f) = \max_{D,D' : |D \Delta D'| = 1} \|f(D) - f(D')\|_1$$

where $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_i |x_i - y_i|$

Generalized Sensitivity

- q_1 = # Males with BMI < 25
- q_2 = # Males with BMI > 25
- q = # Males with BMI

- Let f_1 be a function that answers both q_1, q_2
- Let f_2 be a function that answers both q_1, q

- Sensitivity of f_1 = 1
- Sensitivity of f_2 = 2

- An alternate privacy proof for Alg 2 is to show that the generalized sensitivity of $\widetilde{q_1}$, $\widetilde{q_2}$, $\widetilde{q_3}$, $\widetilde{q_4}$ is 1.

Outline

- Recap
  - Laplace Mechanism

- Composition Theorems

- Optimizing accuracy of DP algorithms
  - Utilizing Parallel Composition
  - Postprocessing & Inference
  - Strategy Selection
  - Data dependent noise

Improving utility of Alg 2

Compute:

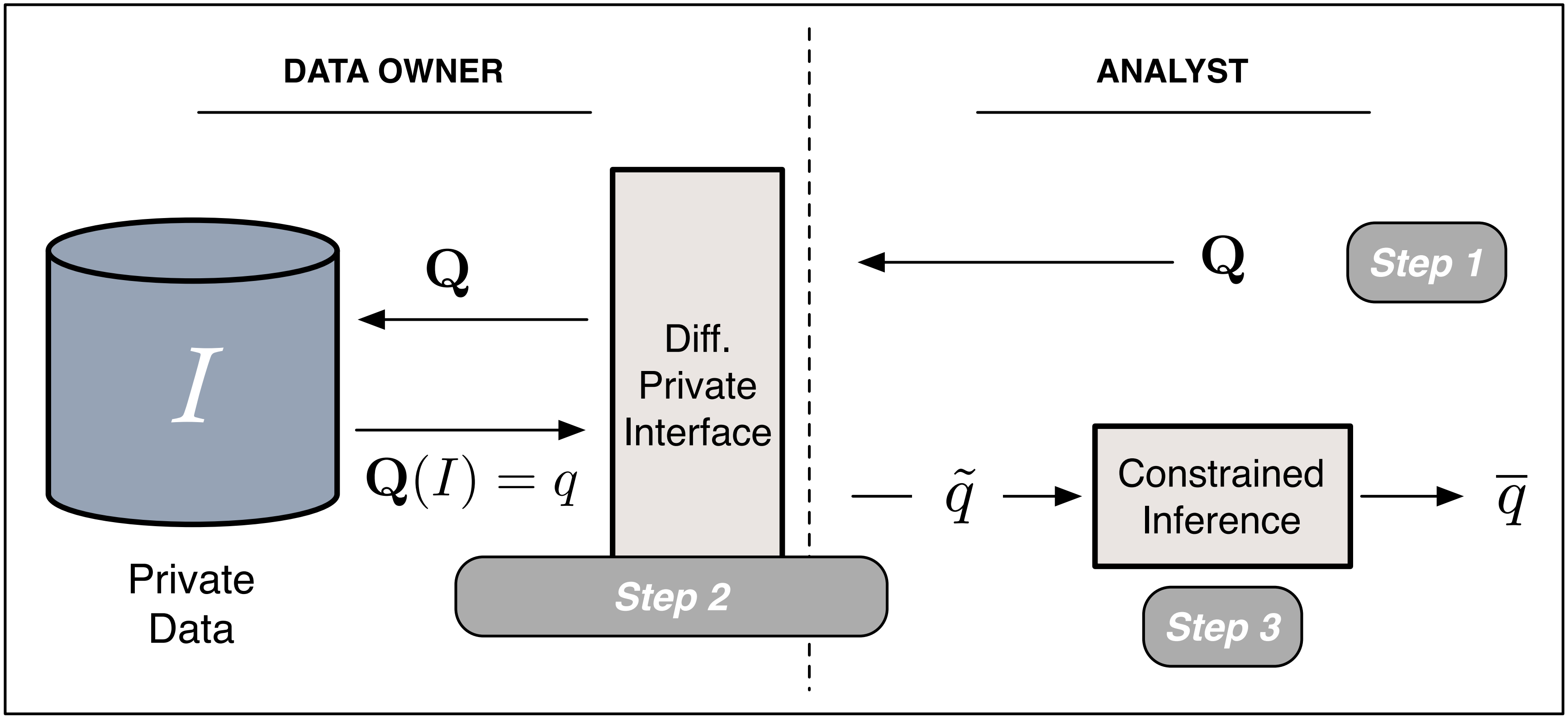
- $\widetilde{q_1}$ = # Males with BMI < 25 + Lap(1/ε)
- $\widetilde{q_2}$ = # Males with BMI > 25 + Lap(1/ε)

Return

- $\widetilde{q_1}$, $\widetilde{q_1}+\widetilde{q_2}$

We know $q_1 \leq q_1 + q_2$,
but P[$\widetilde{q_1} > \widetilde{q_1}+\widetilde{q_2}$] > 0

Constrained Inference



Constrained Inference

- $q_1, q_2, \ldots, q_k$ be a set of queries
- $\widetilde{q_1}, \widetilde{q_2}, \ldots, \widetilde{q_k}$ be the noisy answers
- Constraint $C(q_1, q_2, \ldots, q_k) = 1$ holds on true answers (for all typical databases), but does not hold on noisy answers.

- Goal: Find $\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k}$ that are:
  - Close to $\widetilde{q_1}, \widetilde{q_2}, \ldots, \widetilde{q_k}$
  - Satisfy the constraint $C(\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k})$

Least Squares Optimization

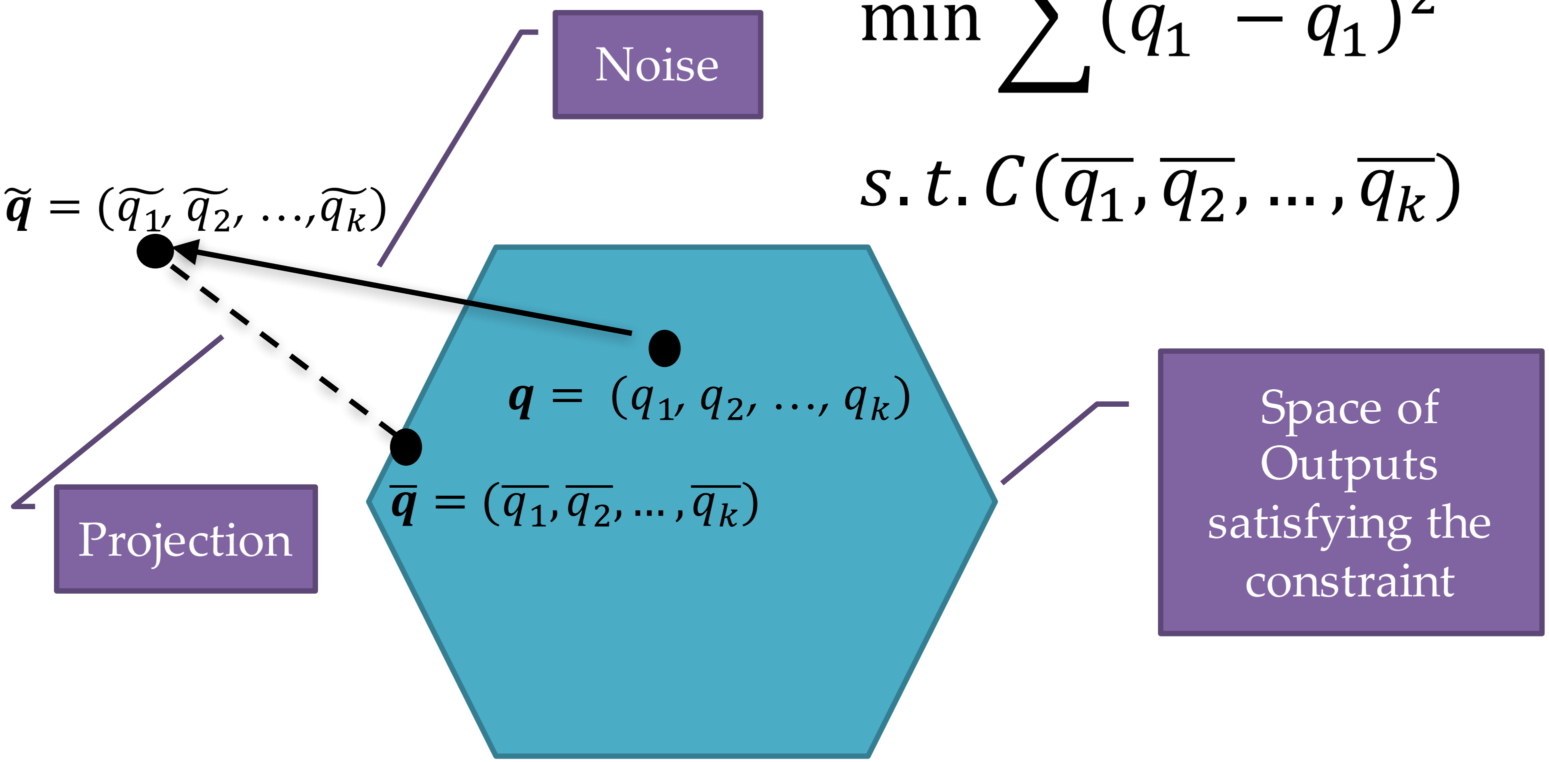
$$\min \sum (\widetilde{q_1} - \overline{q_1})^2$$

$$s.t.\, C(\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k})$$

Geometric Interpretation

$$\min \sum (\widetilde{q_1} - \overline{q_1})^2$$

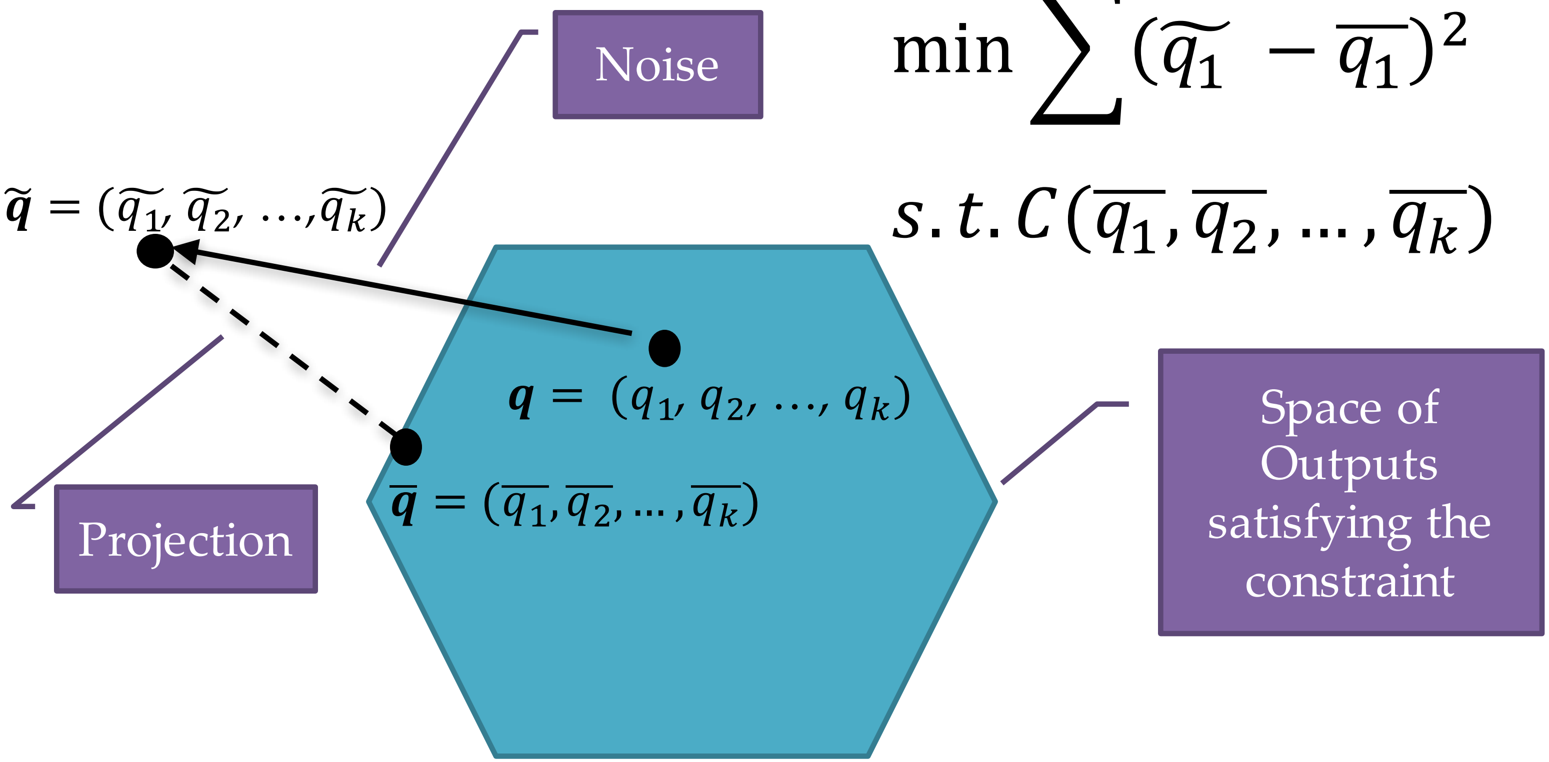
$$s.t.\, C(\overline{q_1}, \overline{q_2}, \dots, \overline{q_k})$$



Geometric Interpretation

$$\min \sum (\widetilde{q_1} - \overline{q_1})^2$$

$$s.t.\, C(\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k})$$



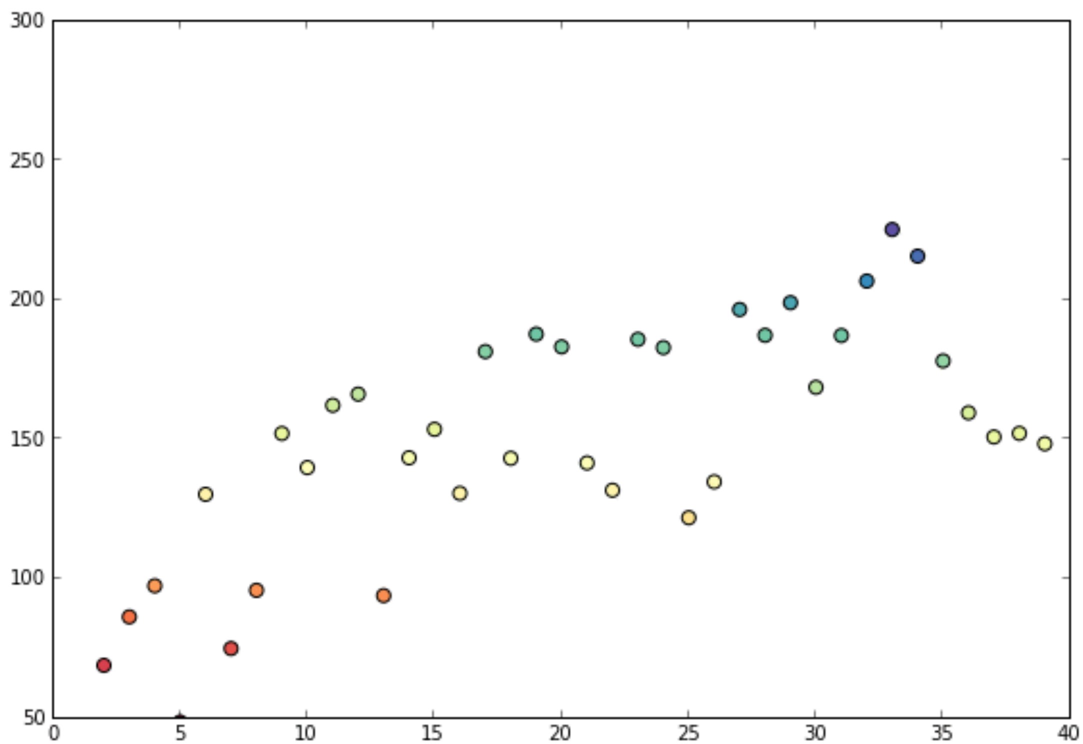
Theorem: $\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_2 \leq \|\boldsymbol{q} - \tilde{\boldsymbol{q}}\|_2$ when the constraints form a convex space

Ordering Constraint

Isotonic Regression:

$$\min \sum (\widetilde{q_1} - \overline{q_1})^2$$

$$s.t. \overline{q_1} \leq \overline{q_1} \leq \ldots \leq \overline{q_k}$$



Outline

- Recap
  - Laplace Mechanism
- Composition Theorems
- Optimizing accuracy of DP algorithms
  - Utilizing Parallel Composition
  - Postprocessing & Inference
  - Strategy Selection
  - Data dependent noise

Problem

Sex	Height	Weight
M	6′2″	210
F	5′3″	190
F	5′9″	160
M	5′3″	180
M	6′7″	250

Queries:

- # people with height in [5′1″, 6′2″]
- # people with height in [2′0″, 4′0″]
- # people with height in [3′3″, 7′0″]
- …

- Design an ε-differentially private algorithm that can answer all range queries.
- What is the total error?

Problem

- Let $\{v_1, \ldots, v_k\}$ be the domain of an attribute
- Let $\{x_1, \ldots, x_k\}$ be the number of rows with values $v_1, \ldots, v_k$

- Range Query: $q_{ij} = x_i + x_{i+1} + \ldots + x_j$

- Goal: Answer all range queries

Strategy 1:

- Answer all range queries using Laplace mechanism

- Sensitivity: O(k^2)
- Total Error: O(k^4/ε^2)

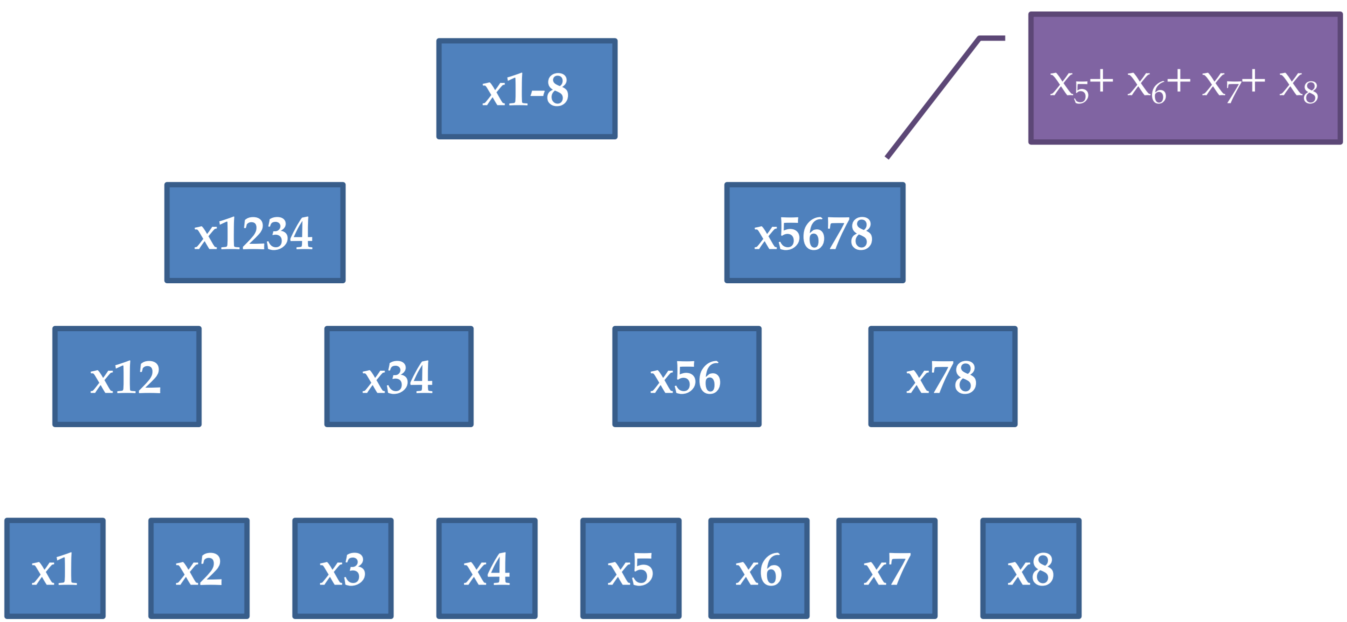
Strategy 2:

- Estimate each individual x_i using Laplace mechanism
- Answer: $q_{ij} = \widetilde{x_i} + \widetilde{x_{i+1}} + \ldots + \widetilde{x_j}$

- Error in each $\widetilde{x_i}$: $O(1/\varepsilon^2)$
- Error in q_{1k}: $O(k/\varepsilon^2)$
- Total Error: $O(k^3/\varepsilon^2)$

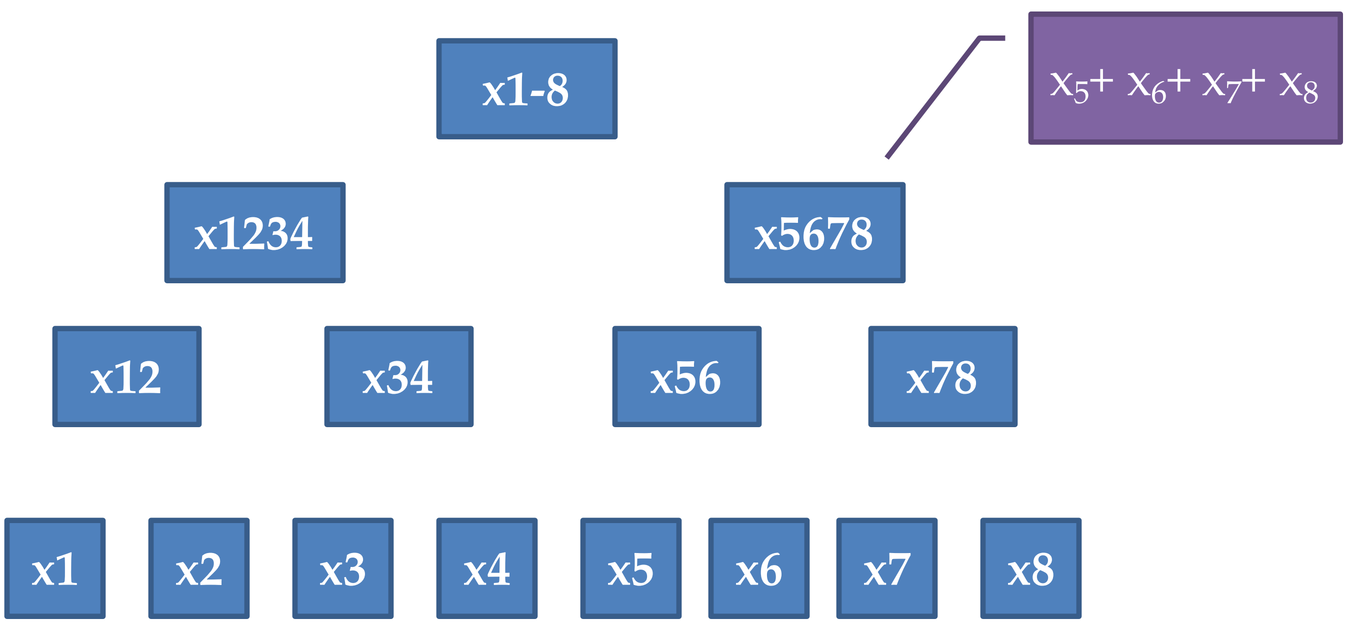
Strategy 3: Hierarchy

- Estimate all the counts in the tree below using Laplace mechanism



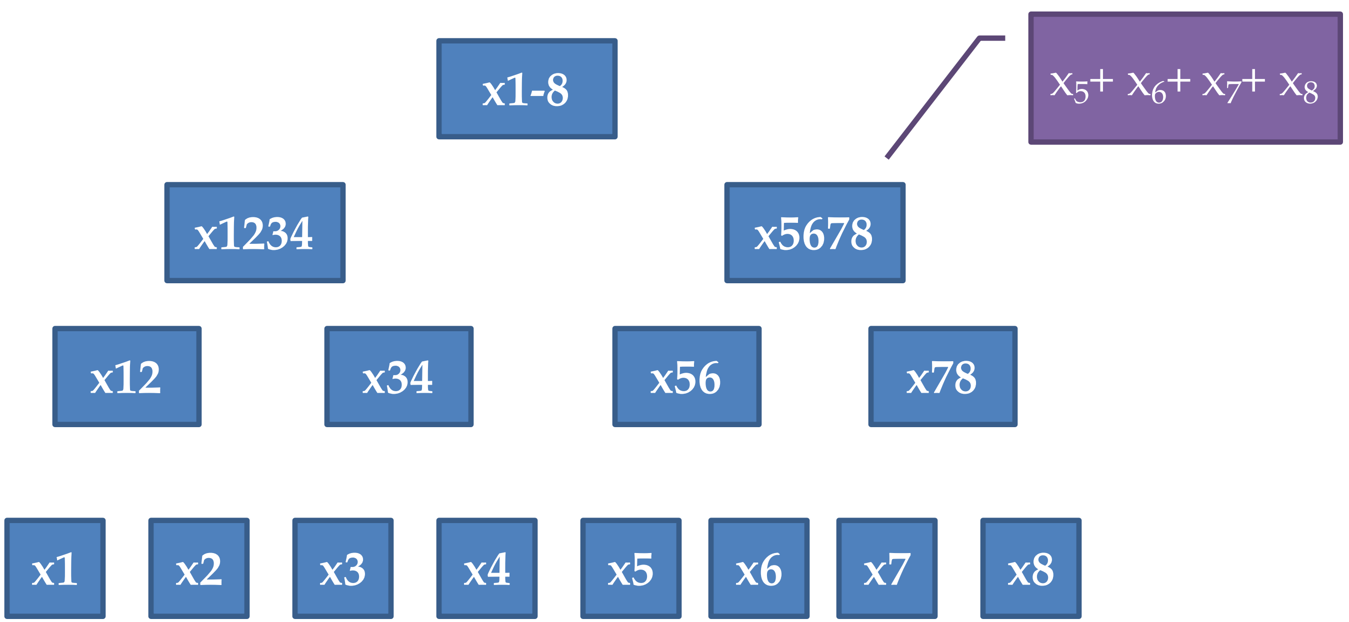
Strategy 3: Hierarchy

- Sensitivity: $\log k$
- Every range query can be answered by summing up at most $2 \log k$ nodes in the tree.



Strategy 3: Hierarchy

- Error in each node: $O((\log k)^2/\varepsilon^2)$
- Max error on a range query: $O((\log k)^3/\varepsilon^2)$
- Total Error: $O(k^2(\log k)^3/\varepsilon^2)$

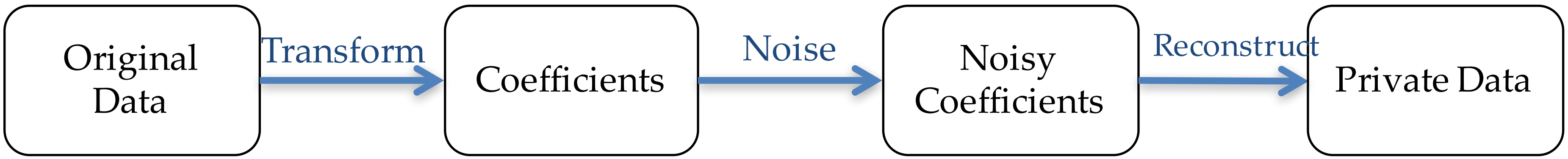


Strategy 3: Hierarchy

- Error in each node: $O((\log k)^2/\varepsilon^2)$
- Max error on a range query: $O((\log k)^3/\varepsilon^2)$
- Total Error: $O(k^2(\log k)^3/\varepsilon^2)$

- Error can be further reduced using constrained inference
  - Here the constraint is that parent counts should not be smaller than child counts.

Strategy based mechanisms

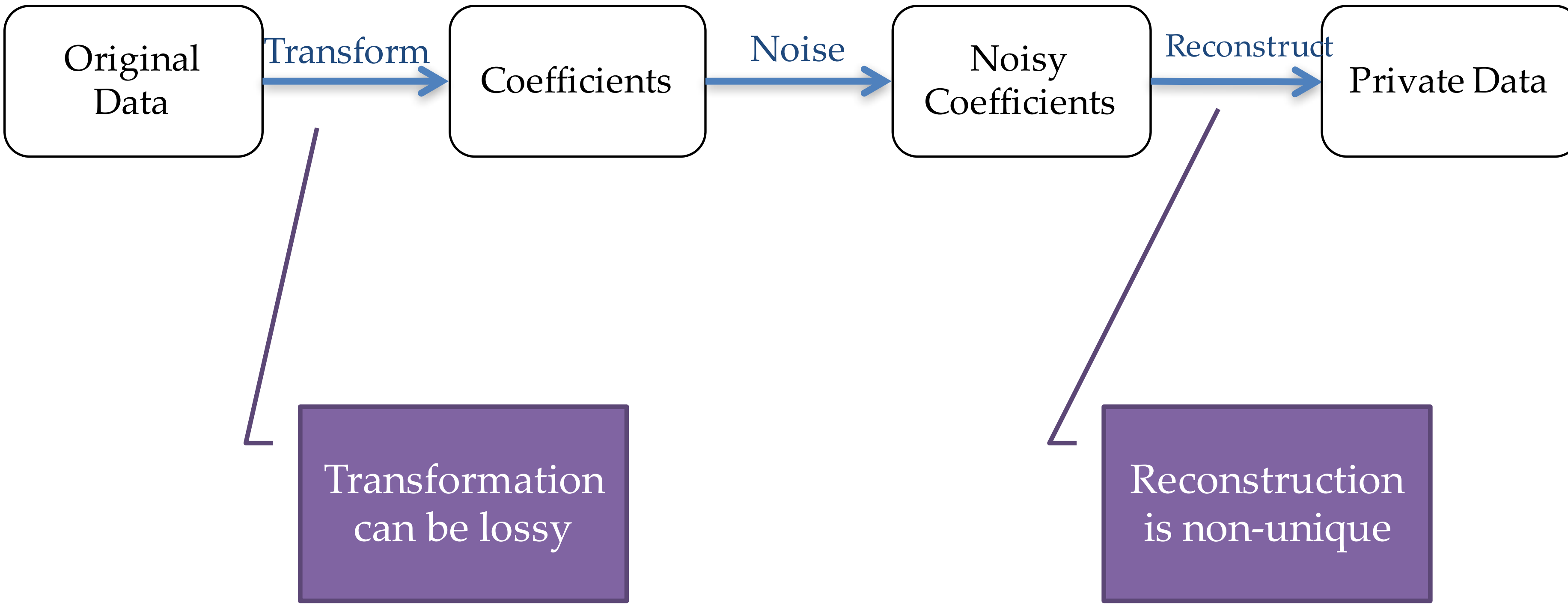


- Can think of nodes in the tree as coefficients.
- Other algorithms use other transformations
  - Wavelets, Fourier coefficients
- Should be able to *losslessly* reconstruct the original data/query answers.
- **General Idea**:
  - Apply transform
  - Add noise to the transformed space (based on sensitivity)
  - Reconstruct original data/query answers from noisy coefficients

Outline

- Recap
  - Laplace Mechanism
- Composition Theorems
- Optimizing accuracy of DP algorithms
  - Utilizing Parallel Composition
  - Postprocessing & Inference
  - Strategy Selection
  - Data dependent noise

Data dependent noise mechanisms



[LHMY14] Li et al. A data- and workload-aware algorithm for range queries under differential privacy. In PVLDB, 2014.

Data dependent noise mechanisms

- Use a data dependent sensitivity measure called Smooth sensitivity.

K. Nissim, S. Raskhodnikova, A. Smith, "Smooth Sensitivity and sampling in private data analysis", STOC 2007

Summary

- Composition theorems help build complex algorithms using simple building blocks
  - Sequential composition
  - Parallel composition
  - Postprocessing
  - *There are more advanced forms of composition.*

Summary

- For the same privacy budget, a better designed algorithm can extract more utility
  - When possible use parallel composition
  - Inference on constraints between queries can reduce error
  - Answering a different *strategy* of queries can help reduce error