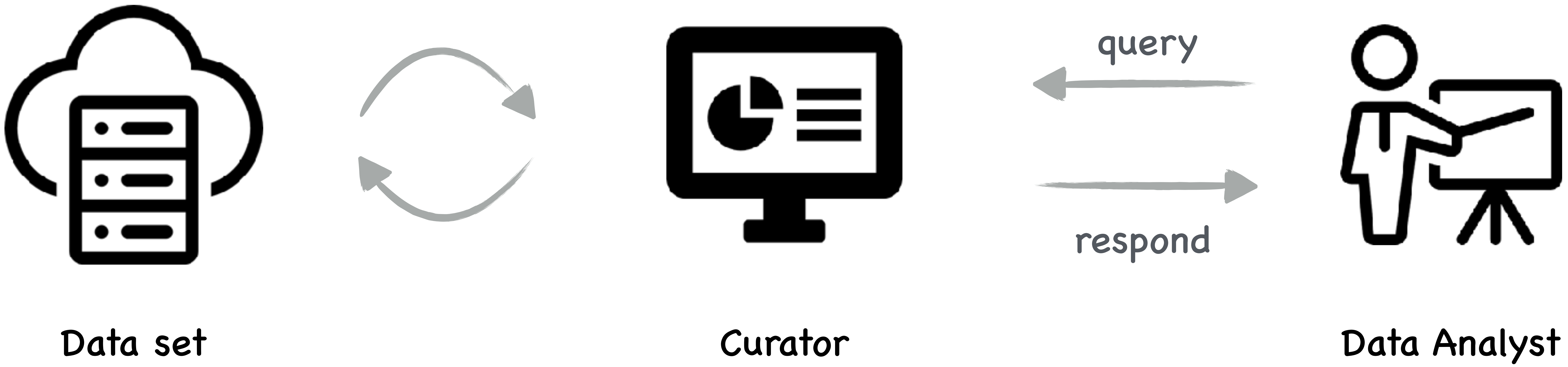
Partial Data Extraction via Noisy Histogram Query: The Information Theoretic Bounds

Wei-Ning Chen, joint work with Prof. I-Hsiang Wang
National Taiwan University

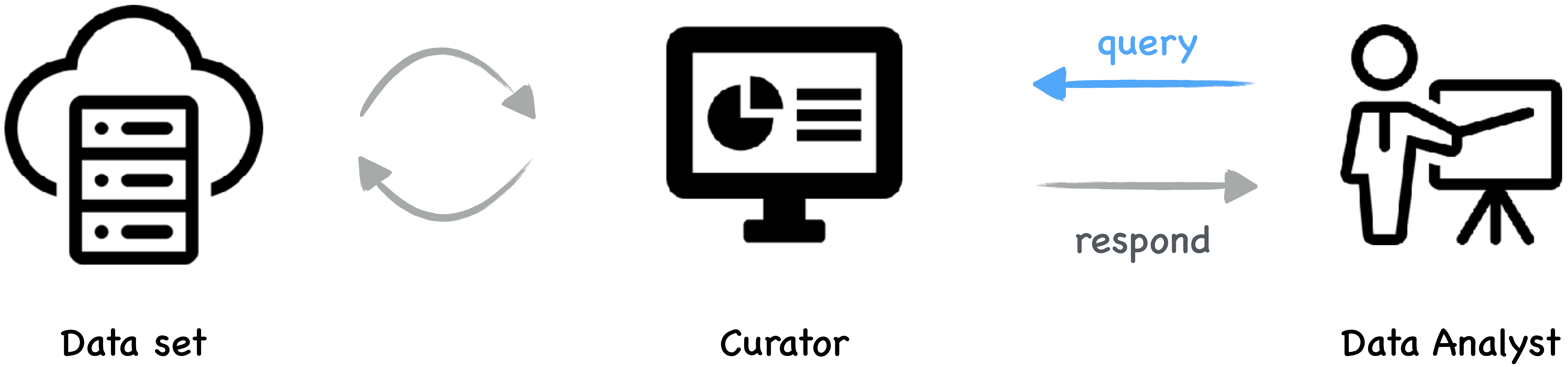
Jun, 2017

Query Model

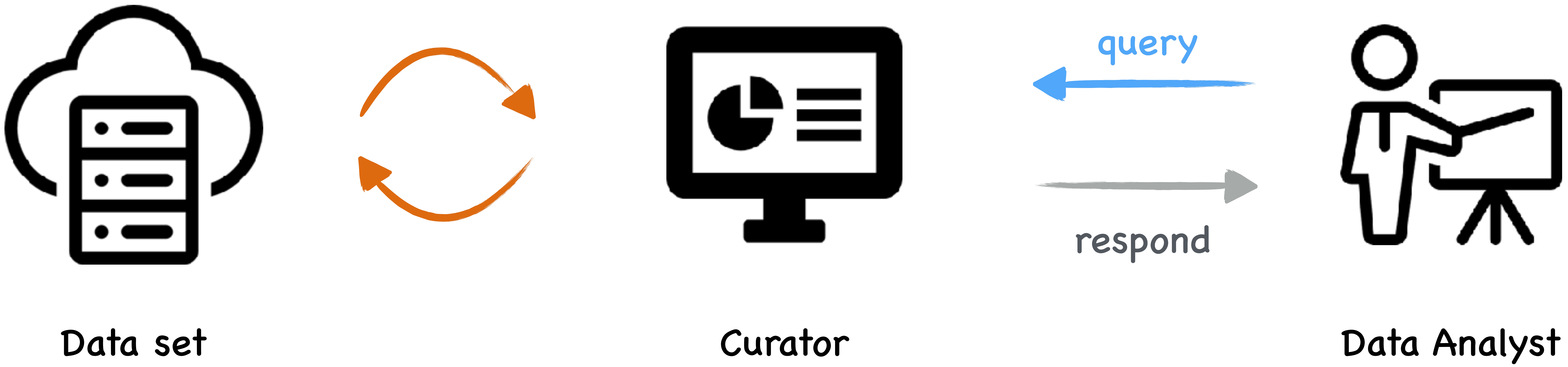


- Query with the curator

Query Model

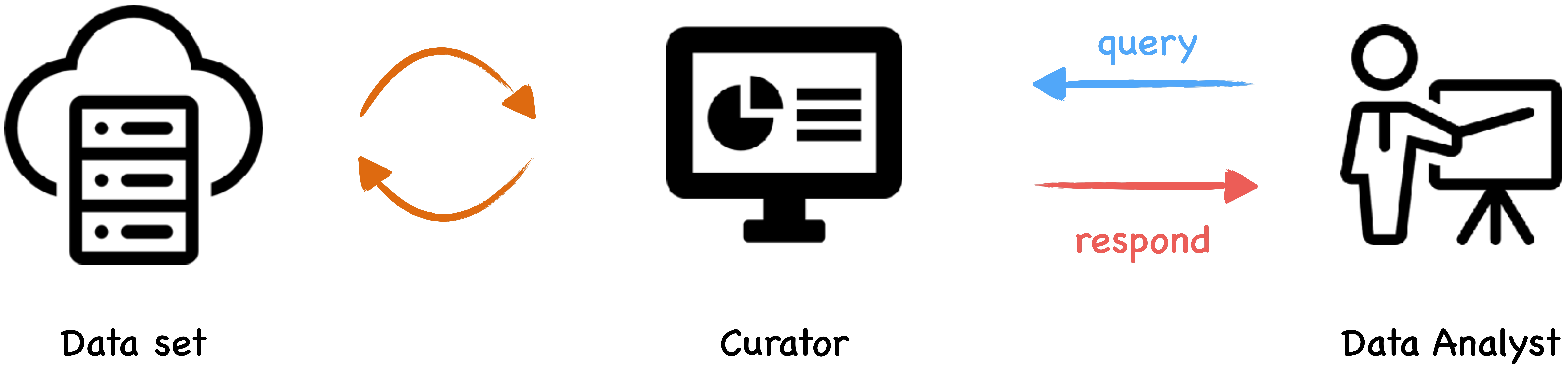


- Query with the curator



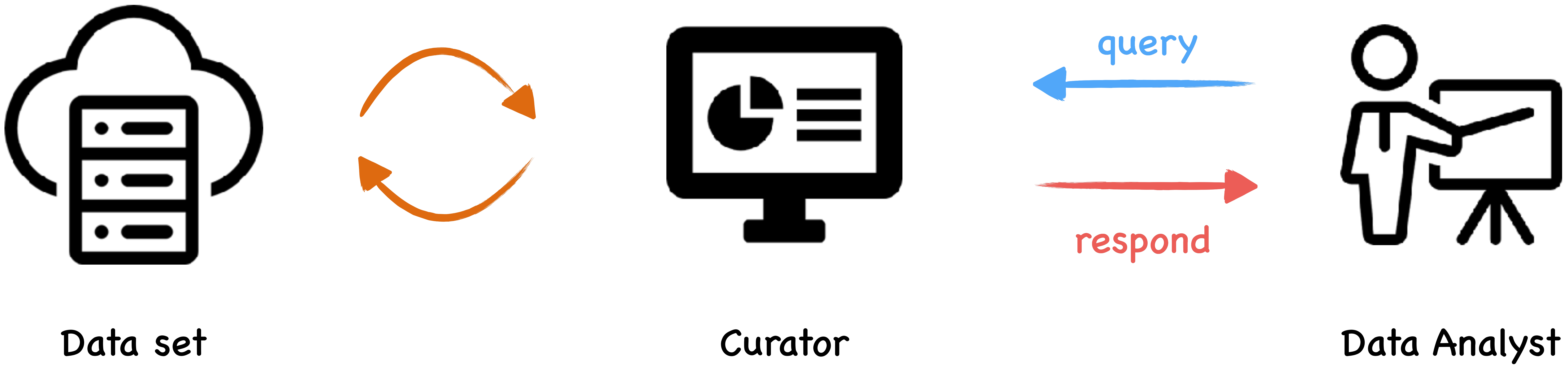
- Query with the curator

Query Model



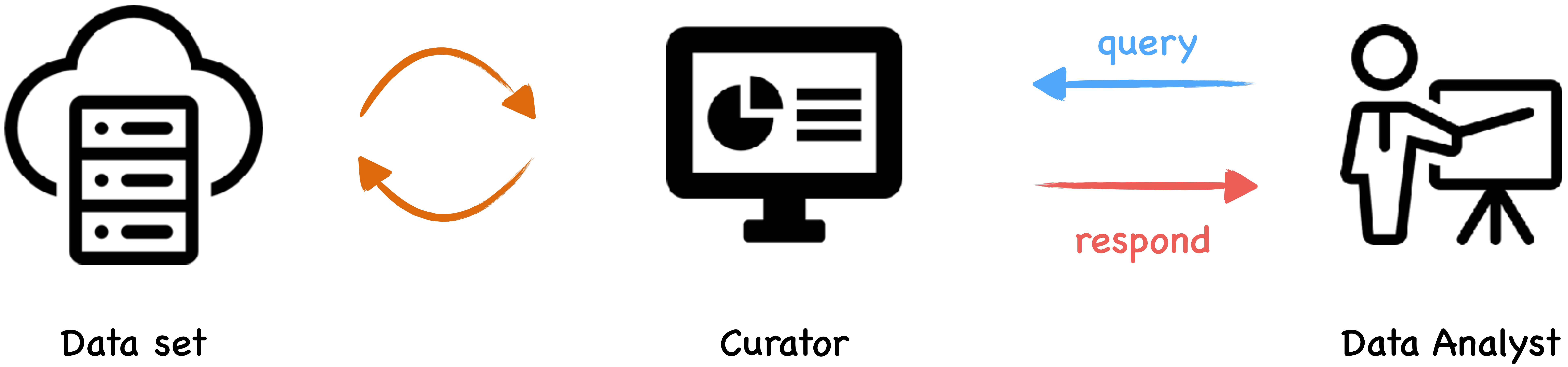
- Query with the curator

Query Model



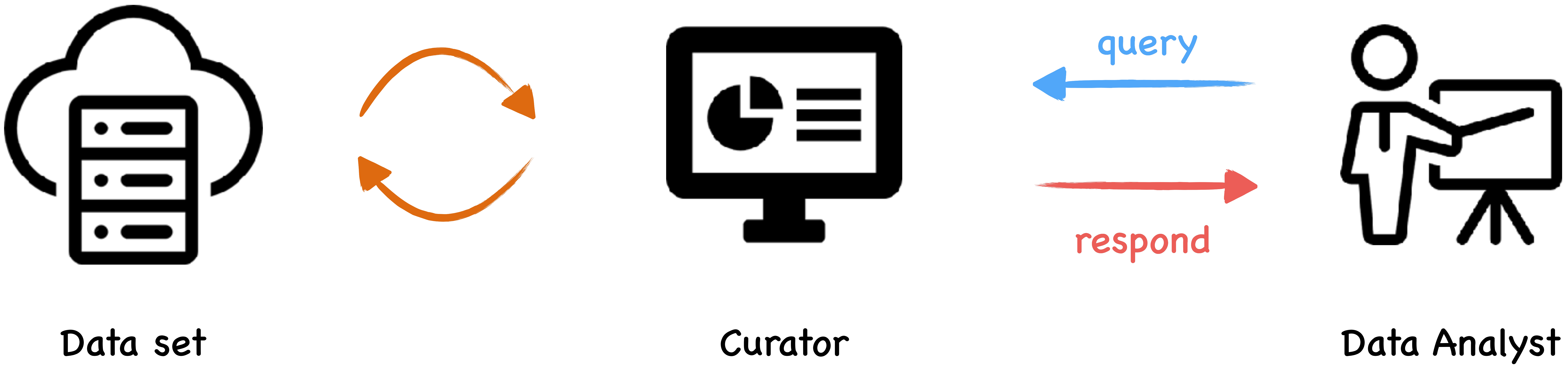
- Query with the curator
- Certain types of queries are allowed

Query Model



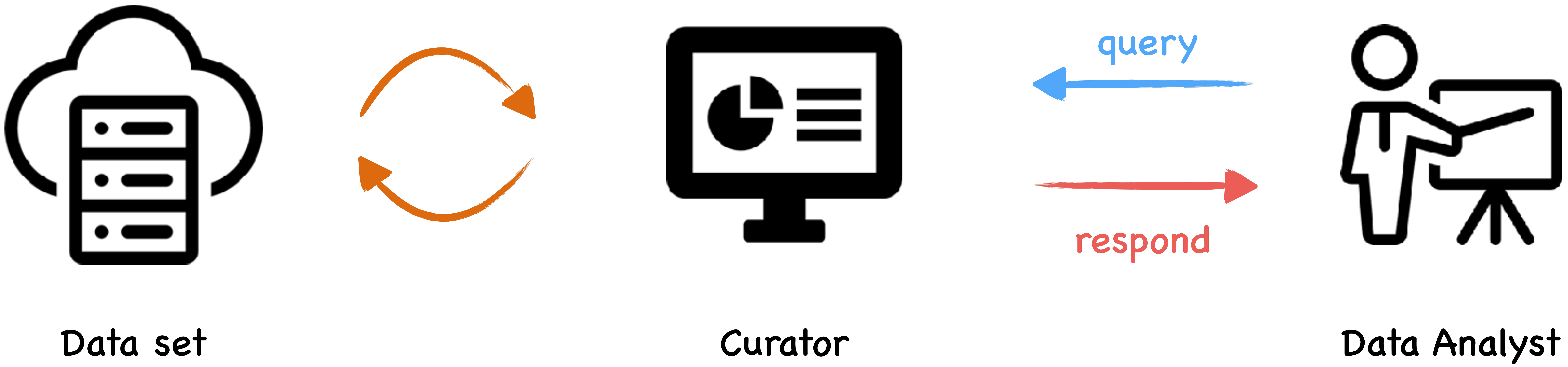
- Query with the curator
- Certain types of queries are allowed
  - Subset query

Query Model



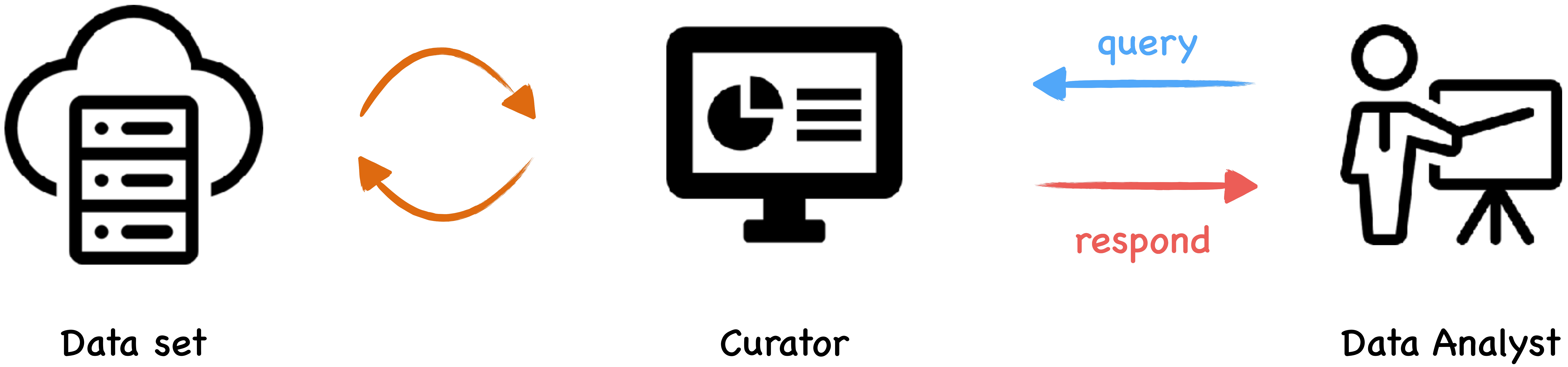
- Query with the curator
- Certain types of queries are allowed
  - Subset query
  - Statistical information of subset

Query Model



- Query with the curator
- Certain types of queries are allowed
  - Subset query
  - Statistical information of subset
- Example :

Query Model

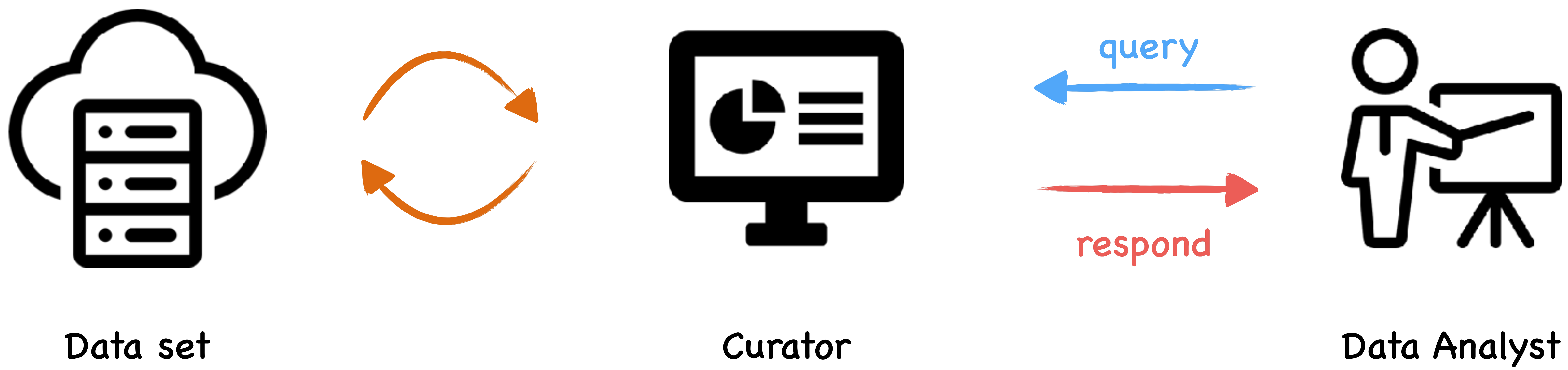


- Query with the curator
- Certain types of queries are allowed
  - Subset query
  - Statistical information of subset
- Example :

  A. Numerical data : statistical mean, variance etc.

  B. Categorical data : counting number, histogram etc.

Query Model



- Query with the curator
- Certain types of queries are allowed
  - Subset quer
  - Statistical information of subset
- Example :

  A. ~~Numerical data : statistical mean, variance etc.~~

  B. Categorical data : counting number, histogram etc.

Histogram Query

- Histogram Query

Users	Blood
1	A
2	A
3	B
4	AB
5	O
6	O

Histogram Query

- Histogram Query

Users	Blood
1	A
2	A
3	B
4	AB
5	O
6	O

User{1,2,3,4}

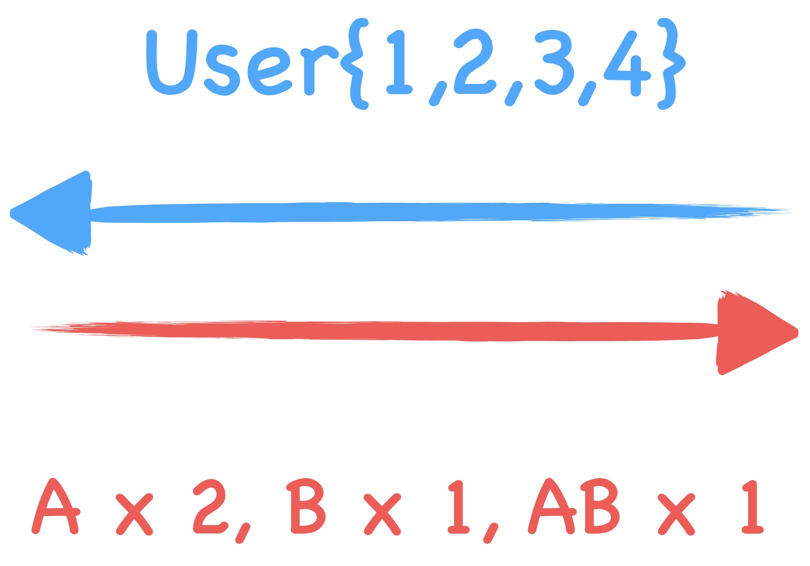
- Histogram Query

Users	Blood
1	A
2	A
3	B
4	AB
5	O
6	O

User{1,2,3,4}

- Histogram Query

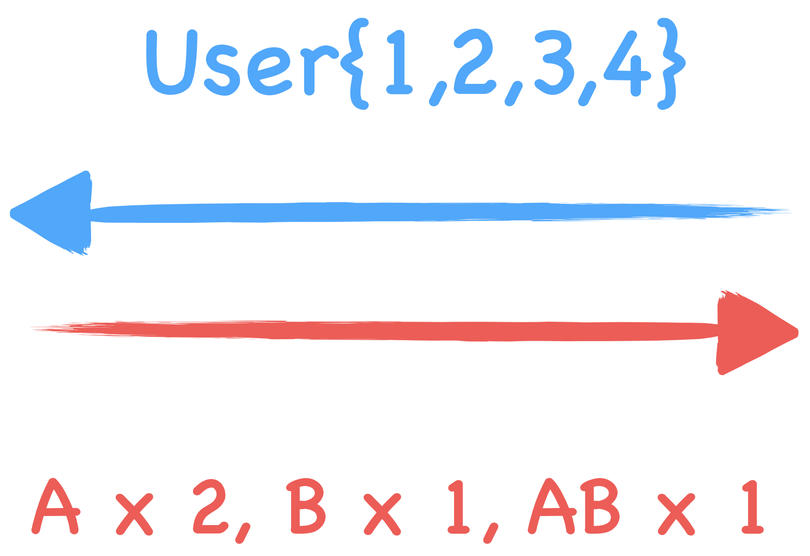
Users	Blood
1	A
2	A
3	B
4	AB
5	O
6	O



Histogram Query

- Histogram Query

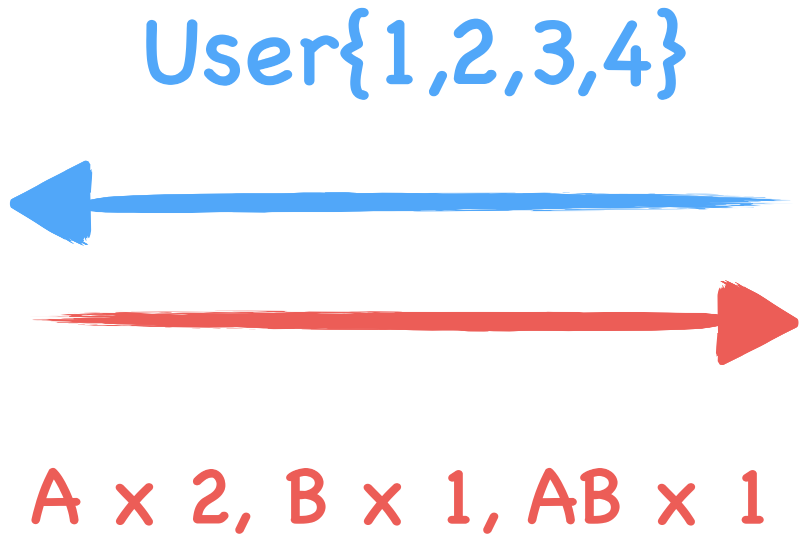
Users	Blood
1	A
2	A
3	B
4	AB
5	O
6	O

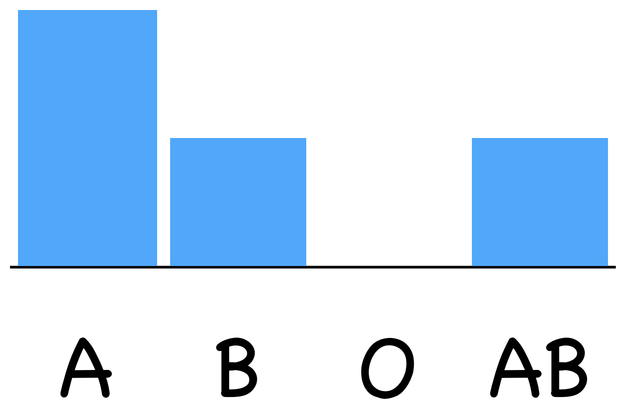


Histogram Query

- Histogram Query

Users	Blood
1	A
2	A
3	B
4	AB
5	O
6	O





The Noisy Response

- Histogram Query

Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O

The Noisy Response

- Histogram Query

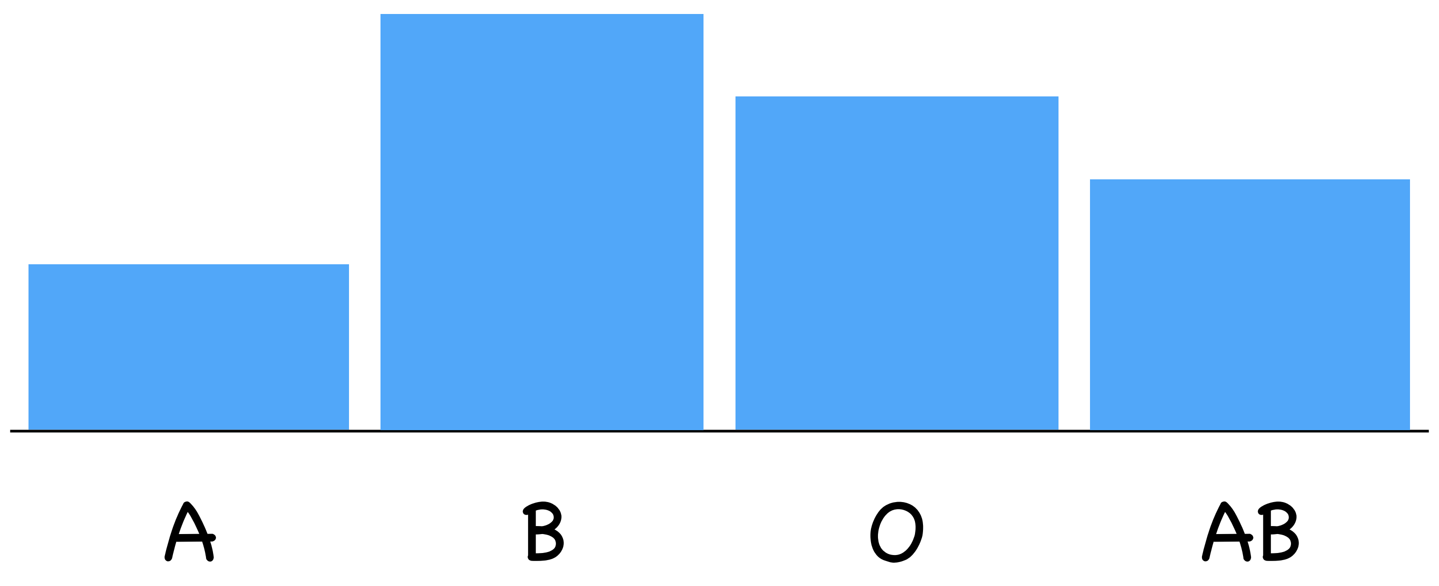
Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O

The Noisy Response

- Histogram Query

Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O

Honest Response

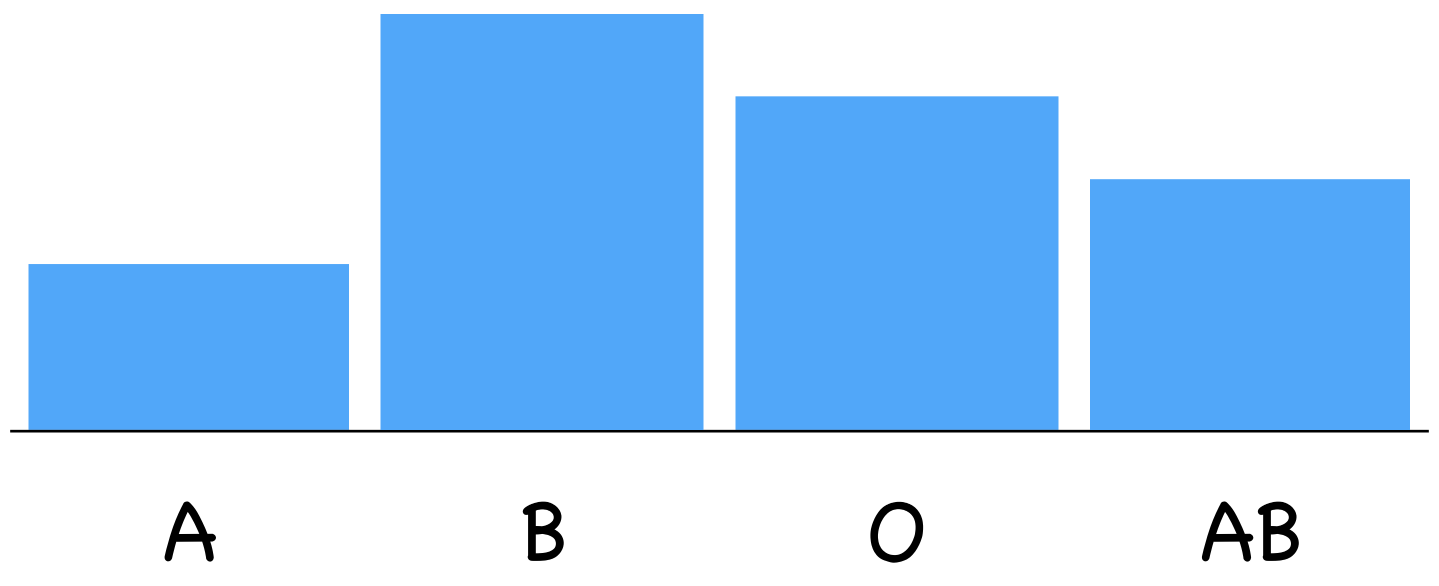


The Noisy Response

- Histogram Query
- Noisy response : ex. to guarantee stronger privacy

Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O

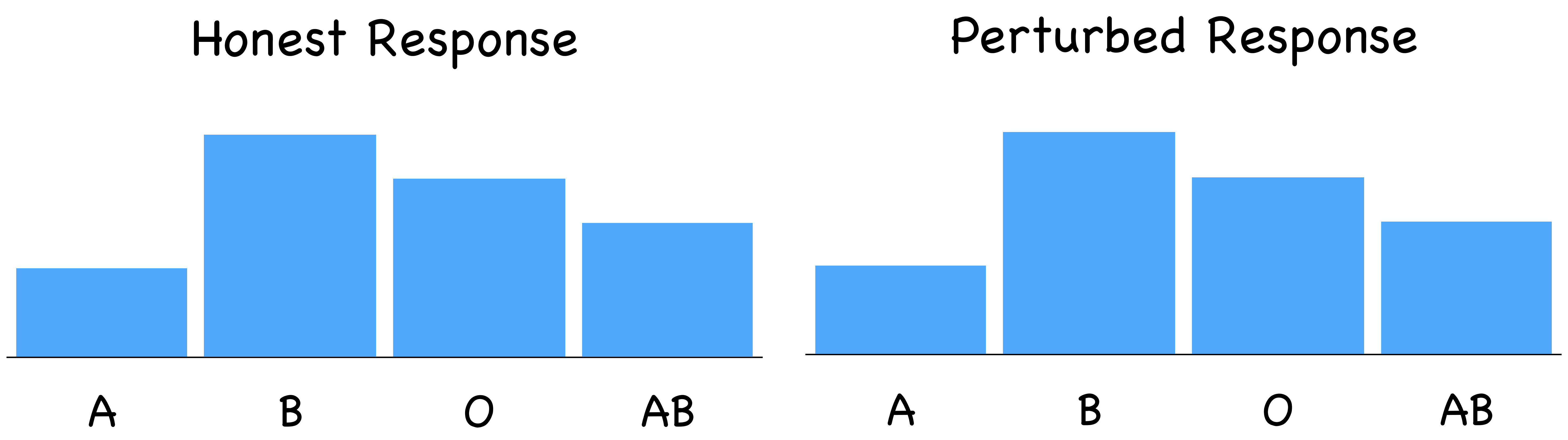
Honest Response



The Noisy Response

- Histogram Query
- Noisy response : ex. to guarantee stronger privacy

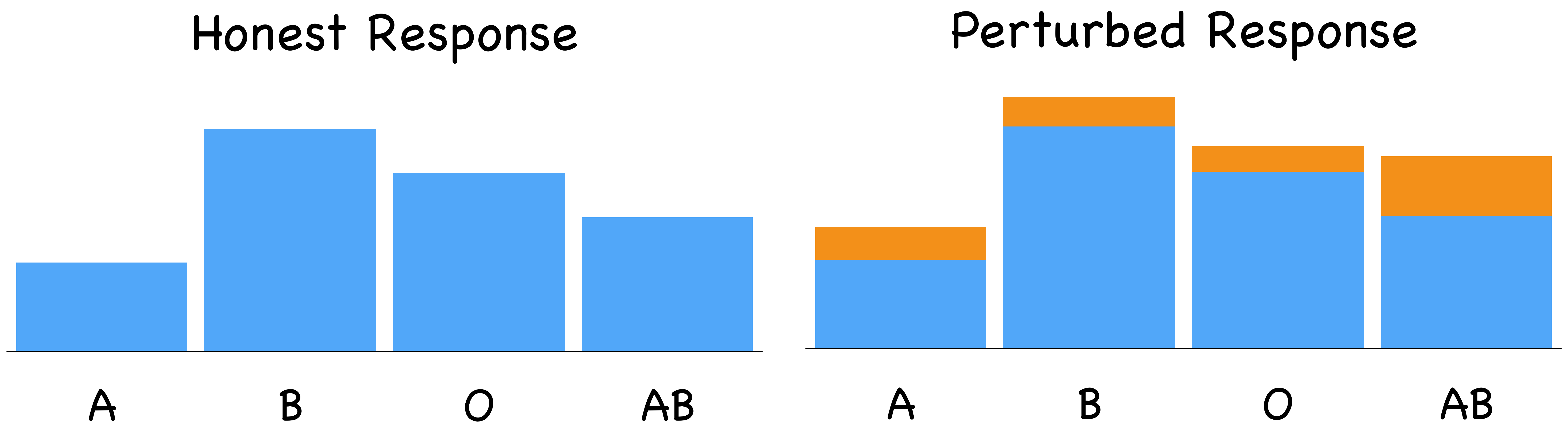
Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O



The Noisy Response

- Histogram Query
- Noisy response : ex. to guarantee stronger privacy

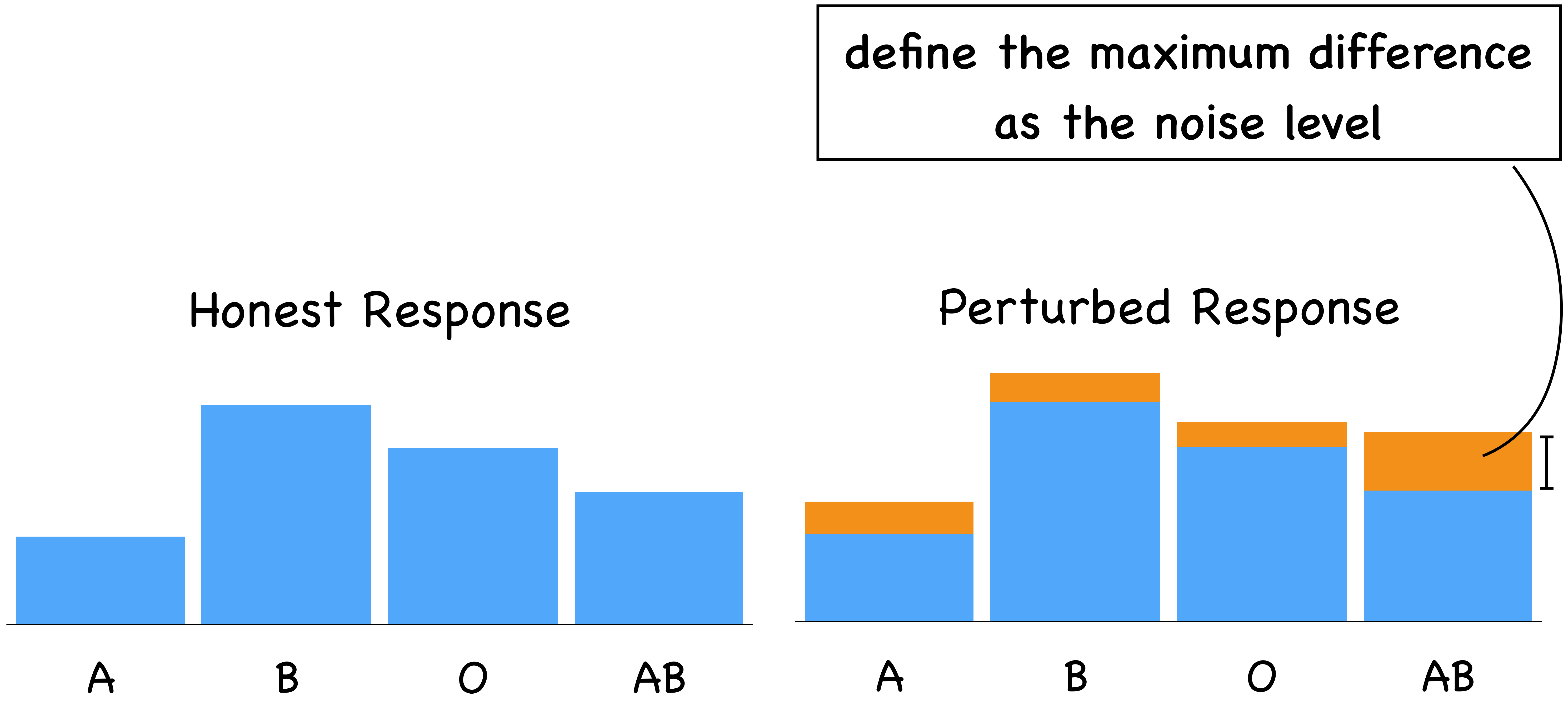
Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O



The Noisy Response

- Histogram Query
- Noisy response : ex. to guarantee stronger privacy

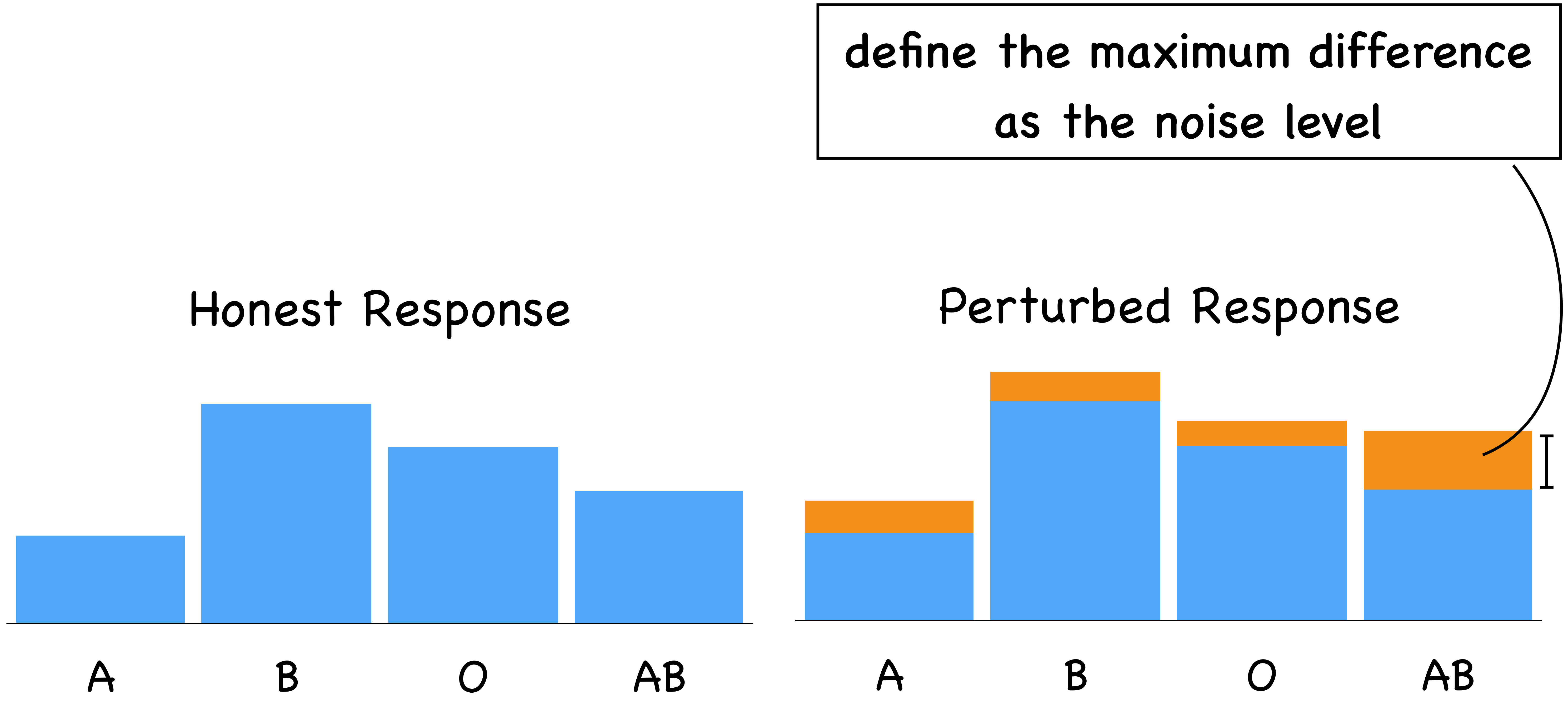
Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O



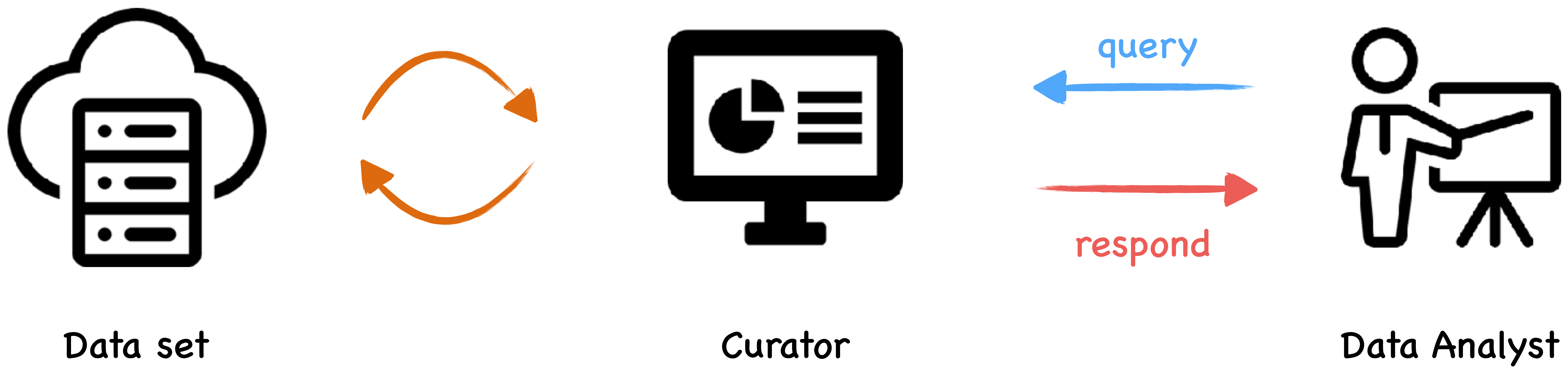
The Noisy Response

- Histogram Query
- Noisy response : ex. to guarantee stronger privacy
- The added noise is at most δ_n

Users	Blood
1	A
2	A
3	B
⋮	⋮
n	O



Problem Statement

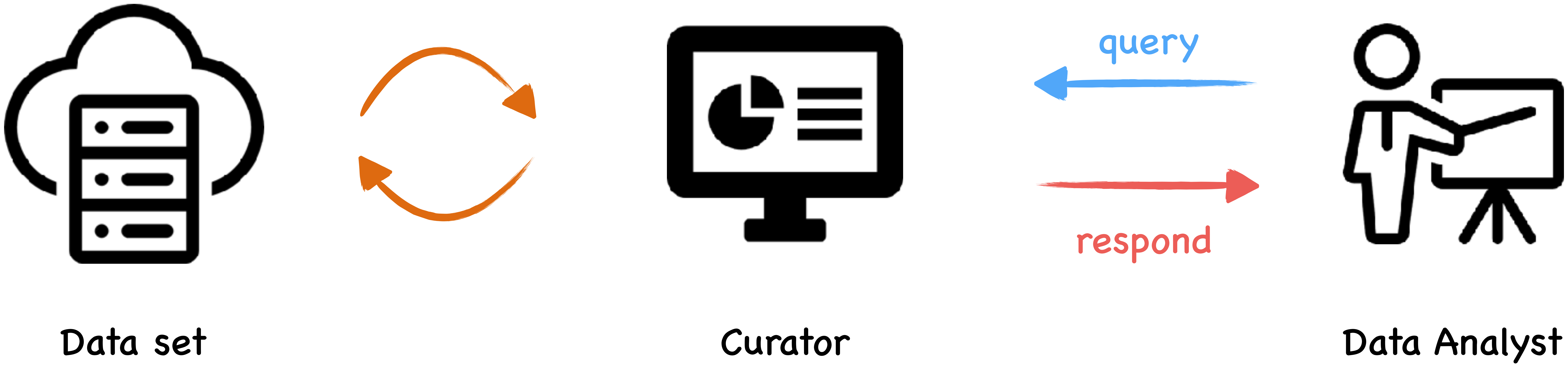


- Goal : to extract the data set partially

[1] I.-H. Wang, et. al "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," Proceedings of IEEE International Symposium on Information Theory 2016

[2] Ahmed El Alaoui , et. al "Decoding from Pooled Data: Phase Transitions of Message Passing ," Proceedings of IEEE International Symposium on Information Theory 2017

Problem Statement

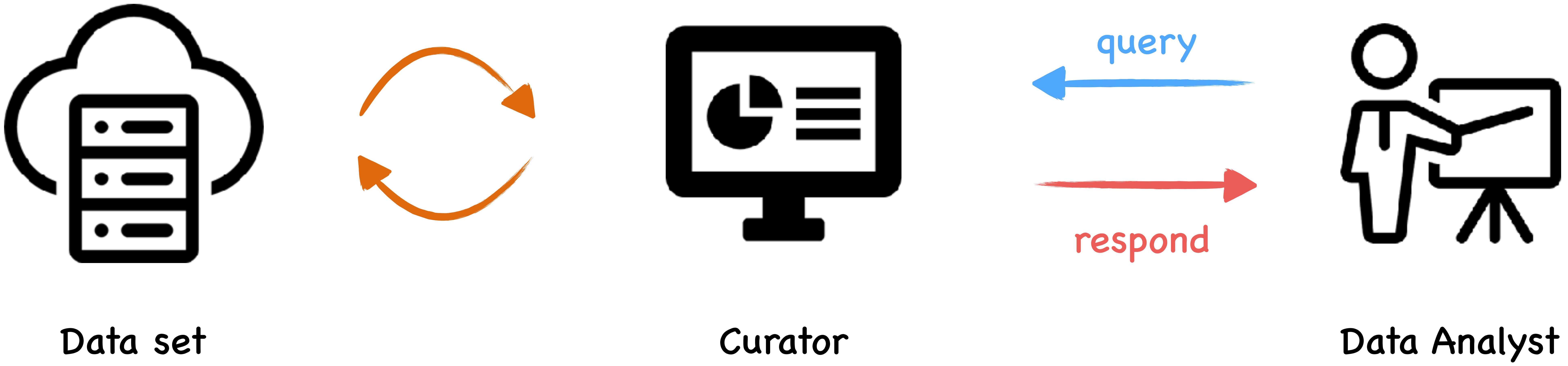


- Goal : to extract the data set partially
  - motivation: privacy, cost of data extraction, etc.

[1] I.-H. Wang, et. al "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," Proceedings of IEEE International Symposium on Information Theory 2016

[2] Ahmed El Alaoui , et. al "Decoding from Pooled Data: Phase Transitions of Message Passing ," Proceedings of IEEE International Symposium on Information Theory 2017

Problem Statement

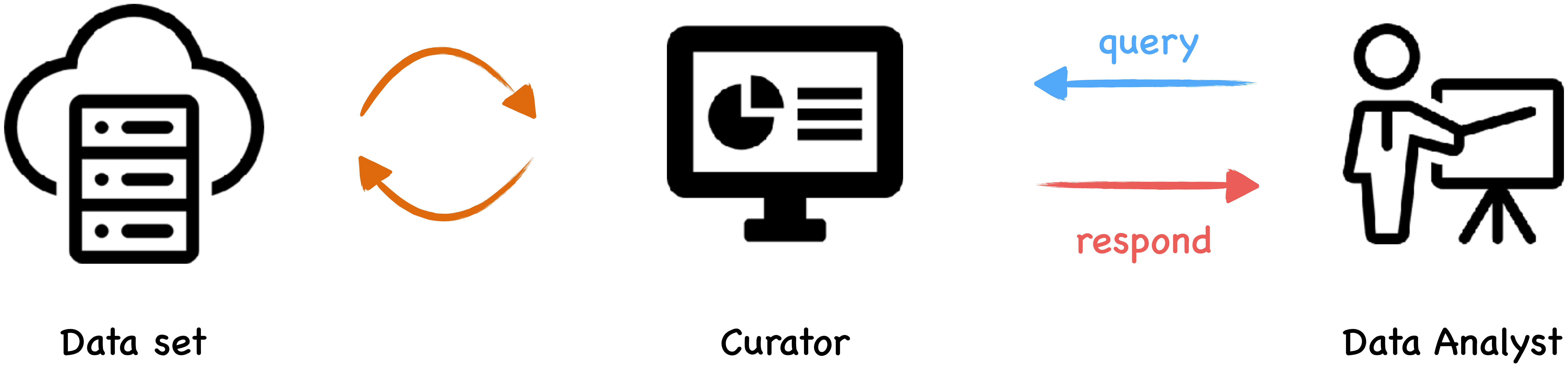


- Goal : to extract the data set partially
  - motivation: privacy, cost of data extraction, etc.
- Key question : how many queries does the analyst required ?

[1] I.-H. Wang, et. al "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," Proceedings of IEEE International Symposium on Information Theory 2016

[2] Ahmed El Alaoui , et. al "Decoding from Pooled Data: Phase Transitions of Message Passing ," Proceedings of IEEE International Symposium on Information Theory 2017

Problem Statement

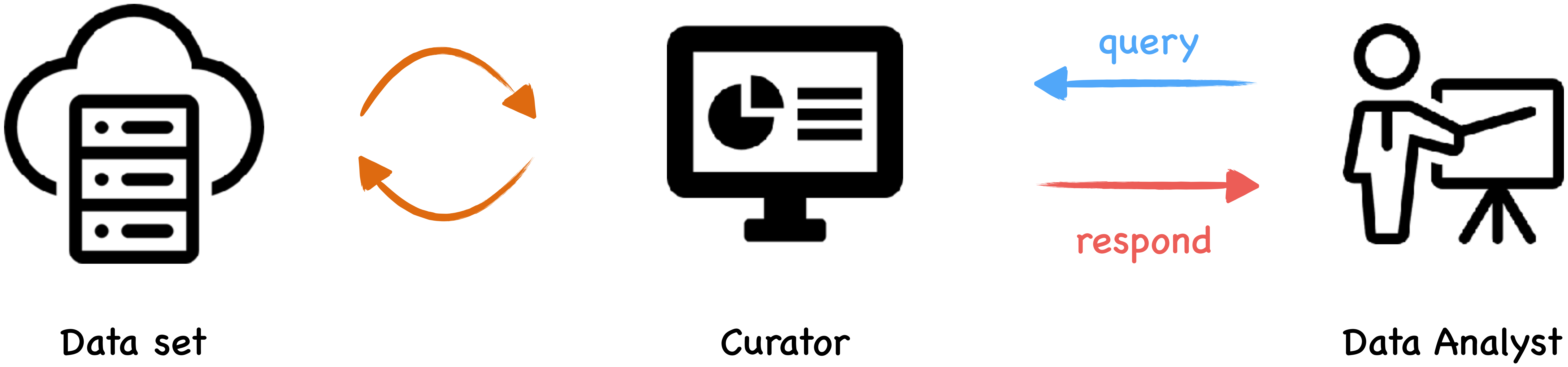


- Goal : to extract the data set partially
  - motivation: privacy, cost of data extraction, etc.
- Key question : how many queries does the analyst required ?
- Query complexity : minimum number of queries to reconstruct the data set

[1] I.-H. Wang, et. al "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," Proceedings of IEEE International Symposium on Information Theory 2016

[2] Ahmed El Alaoui , et. al "Decoding from Pooled Data: Phase Transitions of Message Passing ," Proceedings of IEEE International Symposium on Information Theory 2017

Problem Statement



- Goal : to extract the data set partially
  - motivation: privacy, cost of data extraction, etc.
- Key question : how many queries does the analyst required ?
- Query complexity : minimum number of queries to reconstruct the data set
- In noiseless case, i.e. $\delta_n = 0$, the query complexity in [1] is proven to be $\Theta\left(\frac{n}{\log n}\right)$
  Also, in [2], an AMP algorithm is proposed to decode the data set

[1] I.-H. Wang, et. al "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," Proceedings of IEEE International Symposium on Information Theory 2016
[2] Ahmed El Alaoui , et. al "Decoding from Pooled Data: Phase Transitions of Message Passing ," Proceedings of IEEE International Symposium on Information Theory 2017

Partial Data Reconstruction

- $\mathbf{x}$: original data set

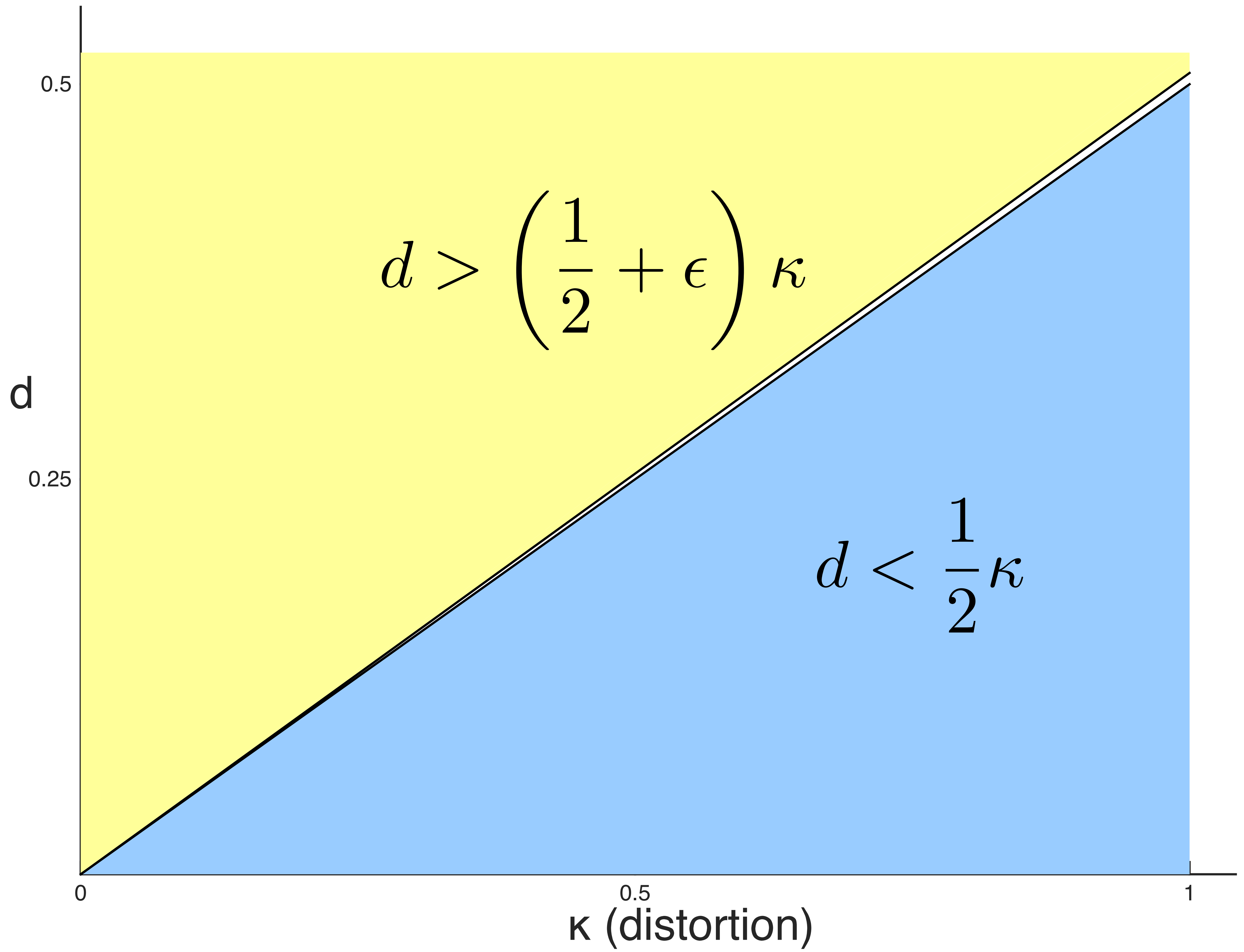
  $\hat{\mathbf{x}}$: recovered data set

- k_n-distortion : $d_{\text{Hamming}}(\mathbf{x}, \hat{\mathbf{x}}) \leq k_n$

$$\mathbf{x} \quad \begin{bmatrix} A \\ B \\ A \\ O \\ AB \\ \vdots \\ O \end{bmatrix} \qquad \hat{\mathbf{x}} \quad \begin{bmatrix} \color{red}{B} \\ B \\ A \\ \color{red}{AB} \\ AB \\ \vdots \\ O \end{bmatrix}$$

Main Result

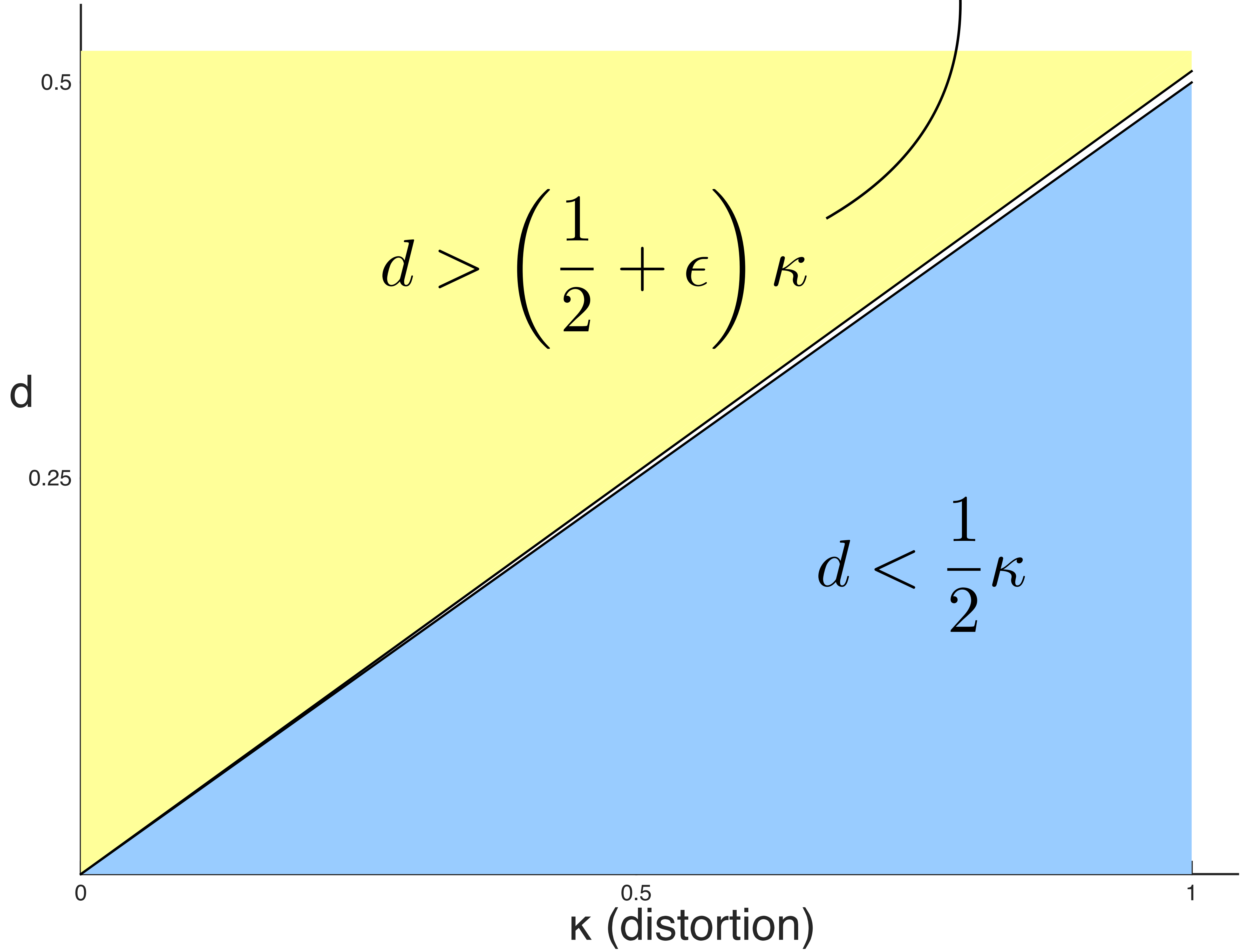
$$\delta_n = \Theta\left(n^d\right), \; k_n = \Theta\left(n^\kappa\right)$$



Main Result

$$\delta_n = \Theta\left(n^d\right),\ k_n = \Theta\left(n^\kappa\right)$$

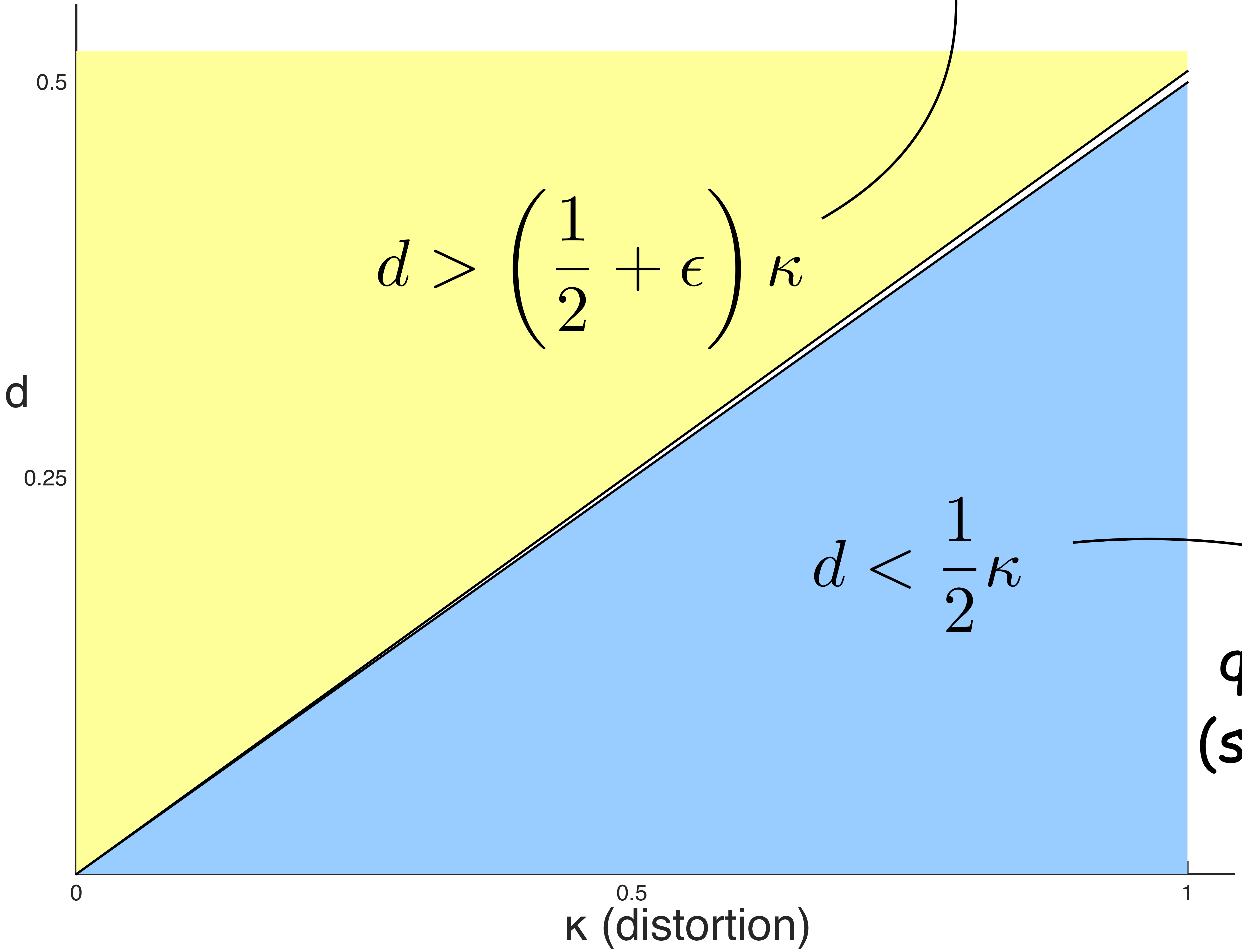
query complexity: non-polynomial



Main Result

$$\delta_n = \Theta\left(n^d\right), \; k_n = \Theta\left(n^\kappa\right)$$

query complexity: non-polynomial

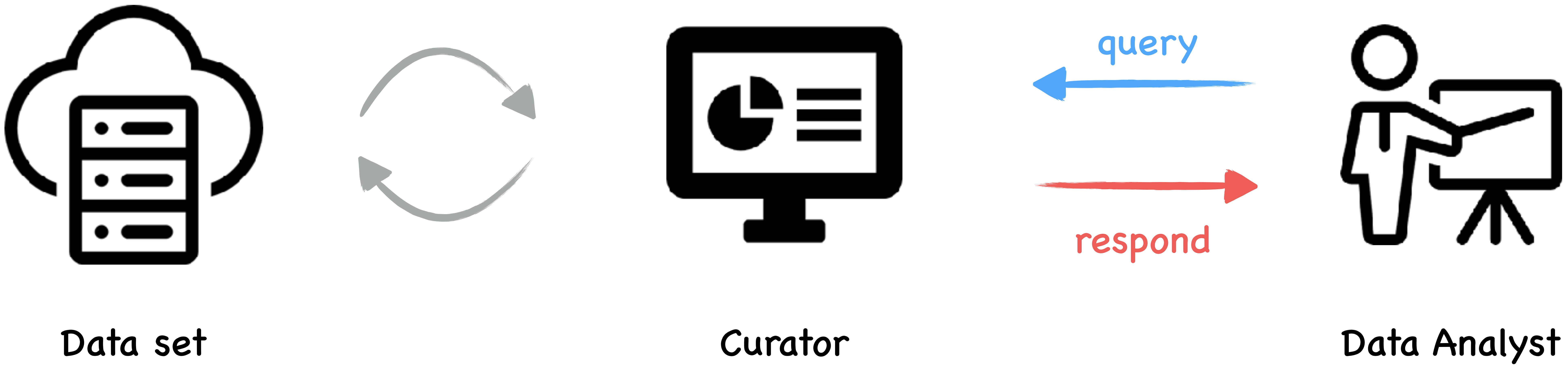


query complexity: sub-linear
(same as the noiseless case !)

The rest of the talk…

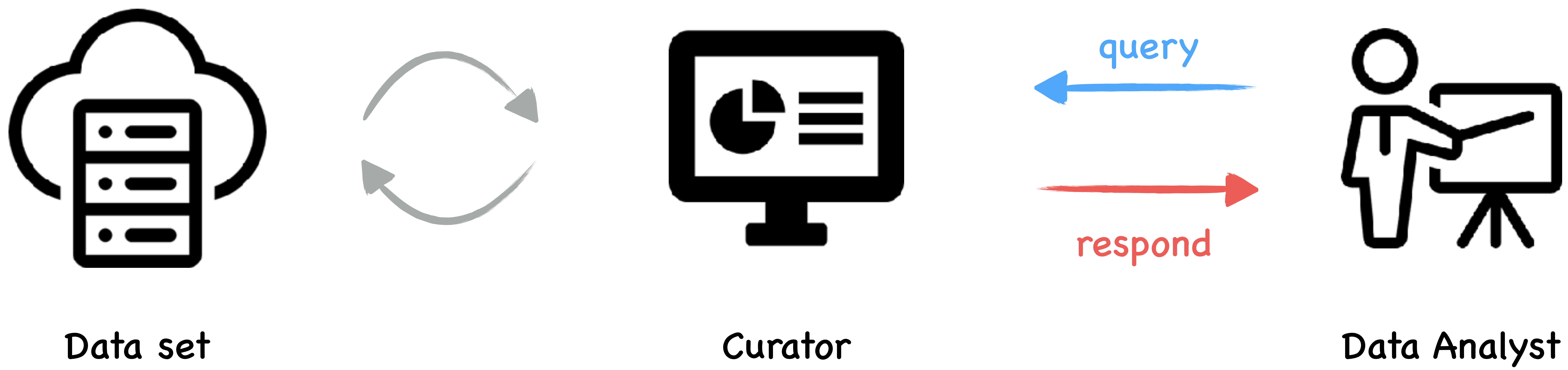
- Problem Formulation
  - Data extraction as a linear inverse problem
- Sketch of Proof :
  A. Regime 1 : Impossibility of Poly-n Query
  B. Regime 2 : The Fundamental Limit of Query Complexity
- Summary

Histogram Query as Linear Multiplication



Users	Blood
1	O
2	B
3	B
⋮	⋮
n	A

Histogram Query as Linear Multiplication

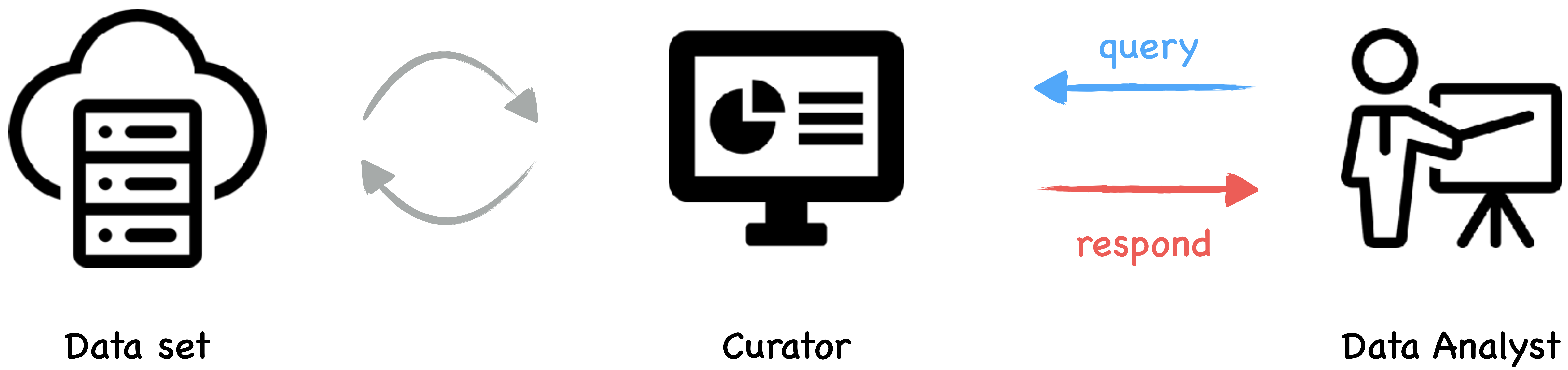


A, B, AB, O

$$n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.$$

$$\mathbf{X}$$

Histogram Query as Linear Multiplication



A, B, AB, O

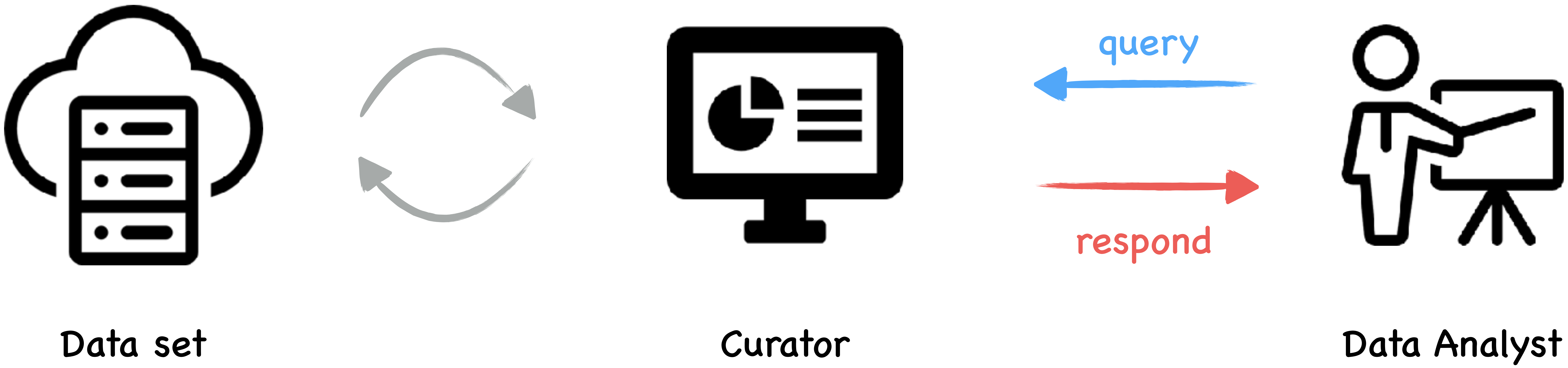
$$n\left\{\begin{bmatrix}0 & 0 & 0 & 1\\ 0 & 1 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots\\ 1 & 0 & 0 & 0\end{bmatrix}\right.$$

$$\mathbf{X}$$

User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix}1 & 1 & 0 & \cdots & 0 & 1\end{bmatrix}}_{n}$$

Histogram Query as Linear Multiplication



A, B, AB, O

$$n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.$$

$$\mathbf{X}$$

User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X}$$

Histogram Query as Linear Multiplication

Data set | Curator | Data Analyst

A, B, AB, O

$$
n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.
$$

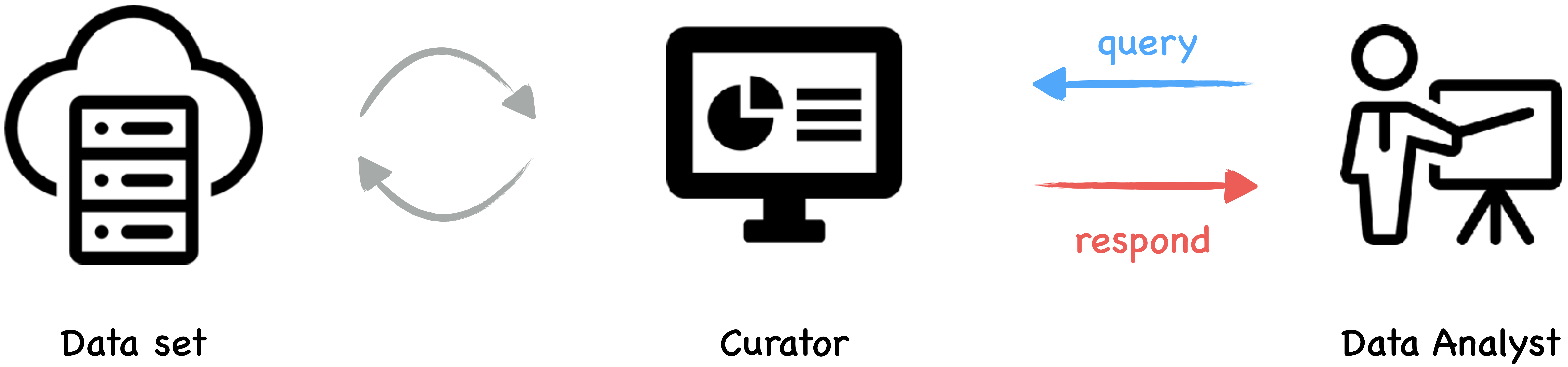
$$\mathbf{X}$$

User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X} + \Delta_i$$

Histogram Query as Linear Multiplication



A, B, AB, O

$$n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.$$

$\mathbf{X}$

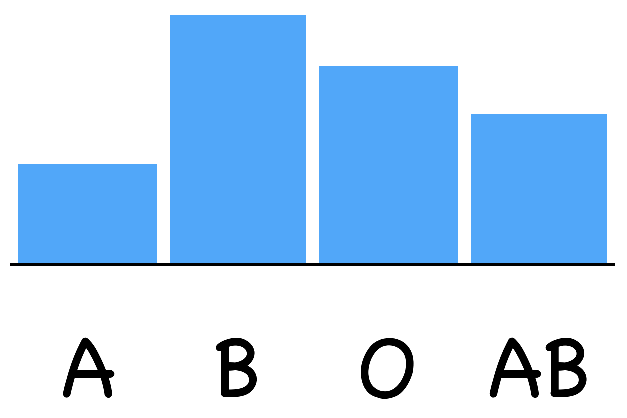
User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

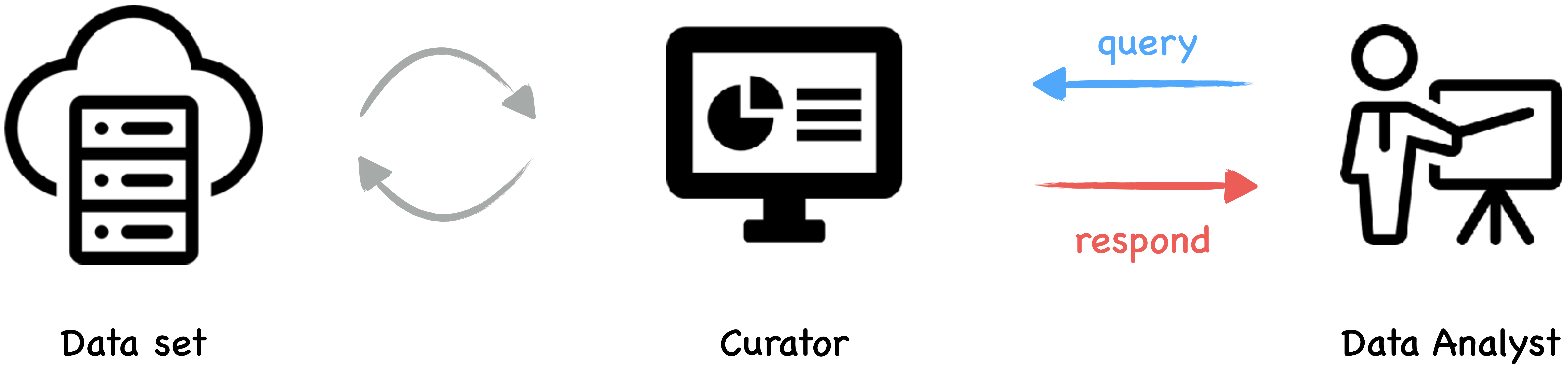
$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X} + \Delta_i$$

A, B, O, AB

$$\mathbf{y}_i = [10, 20, 18, 28]$$



Histogram Query as Linear Multiplication



A, B, AB, O

$$n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.$$

$\mathbf{X}$

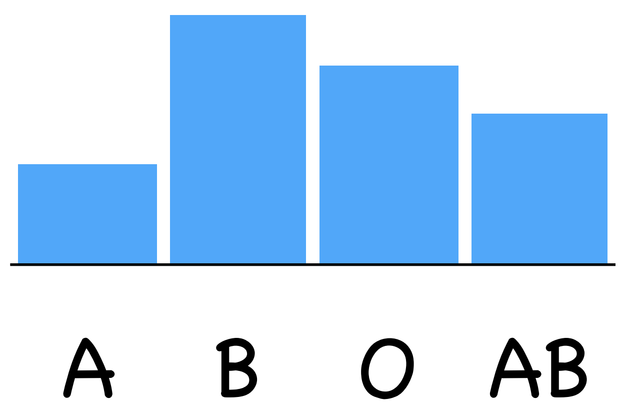
User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X} + \Delta_i$$

A, B, O, AB

$$\mathbf{y}_i = [10, 20, 18, 28]$$



Histogram Query as Linear Multiplication

Data set　　　　Curator　　　　Data Analyst

A, B, AB, O

$$n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.$$

$\mathbf{X}$

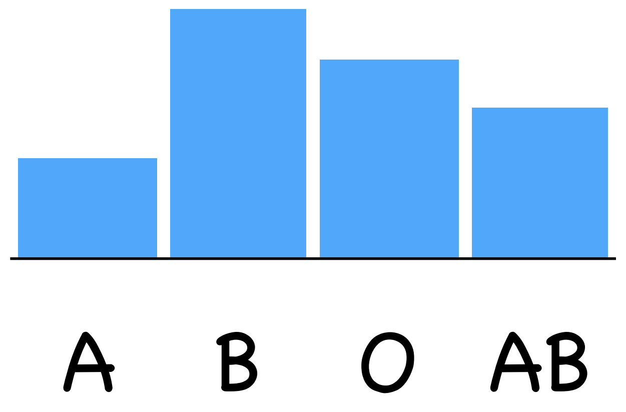
User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X} + \Delta_i$$

A, B, O, AB

$$\mathbf{y}_i = [10, 20, 18, 28]$$



Histogram Query as Linear Multiplication

Data set

Curator

Data Analyst

Decode column by column

A, B, AB, O

$$n \left\{ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix} \right.$$

$\mathbf{X}$

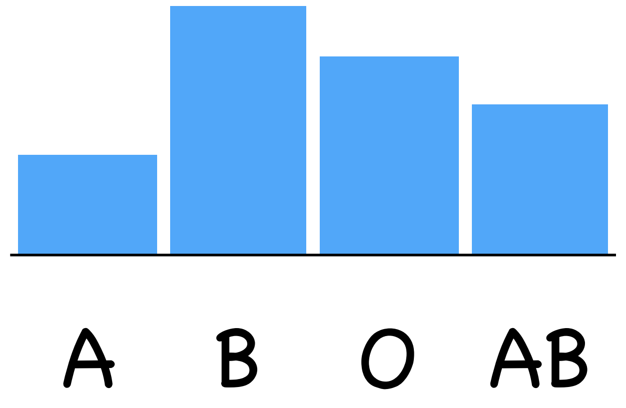
User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

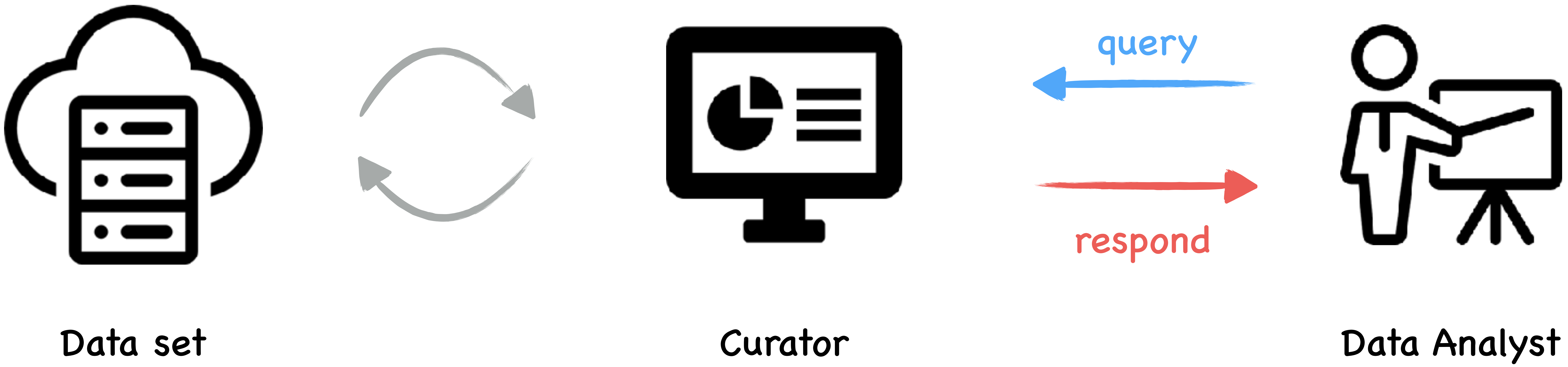
$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X} + \Delta_i$$

A, B, O, AB

$$\mathbf{y}_i = [10, 20, 18, 28]$$



Histogram Query as Linear Multiplication



Decode column by column

A, B, AB, O

$$n\left\{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix}\right.$$

$$\mathbf{X}$$

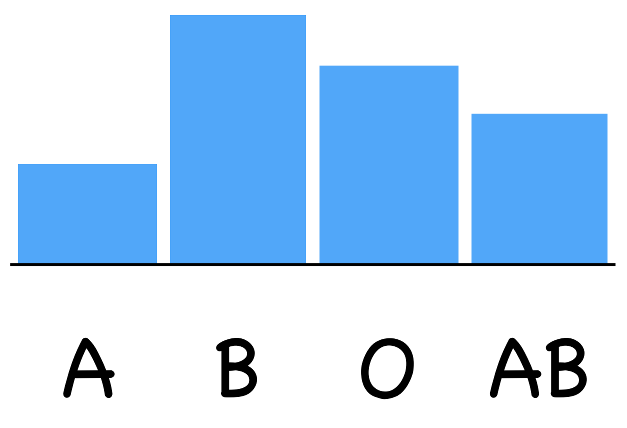
User{1,2,n}

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

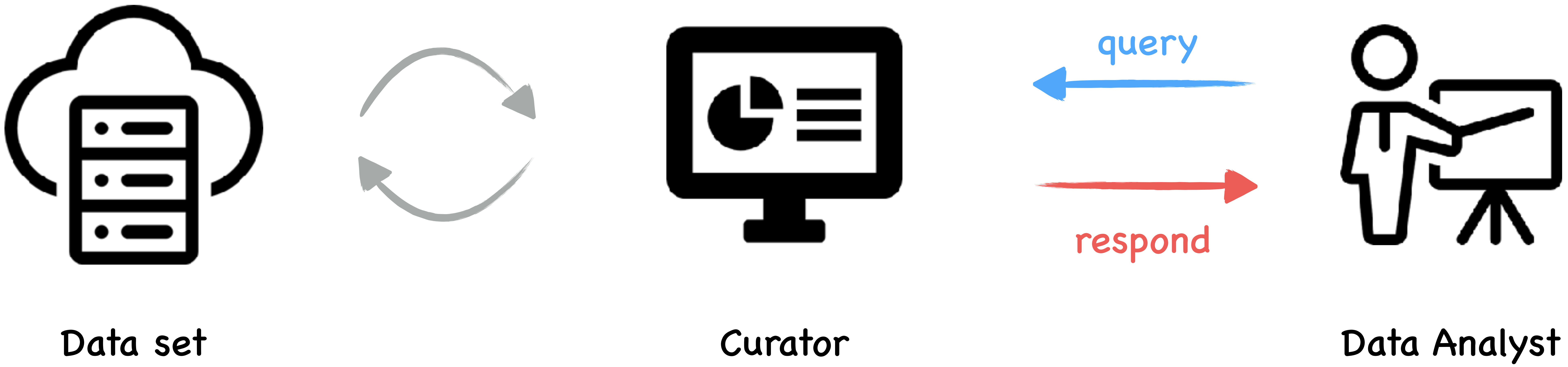
$$\boldsymbol{y}_i = \boldsymbol{q}_i^\intercal \mathbf{X} + \Delta_i$$

A, B, O, AB

$$\mathbf{y}_i = [10, 20, 18, 28]$$



Histogram Query as Linear Multiplication



$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$\boldsymbol{q}_i^\intercal = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix}}_{n}$$

$$y_i = \boldsymbol{q}_i^\intercal \mathbf{x} + \boldsymbol{\Delta}$$

$y_i : \#$ of 1 in $\mathbf{x}$

Histogram Query as Linear Multiplication

Data set | Curator | Data Analyst

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \boldsymbol{q}_1^\intercal \\ \boldsymbol{q}_2^\intercal \\ \vdots \\ \boldsymbol{q}_{T_n}^\intercal \end{bmatrix}$$

$$\boldsymbol{y} = \mathbf{Q}\mathbf{x} + \boldsymbol{\Delta}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{T_n} \end{bmatrix} + \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_{T_n} \end{bmatrix}$$

The Equivalent Linear Inverse Problem

- Equivalent linear inverse problem :

The Equivalent Linear Inverse Problem

- Equivalent linear inverse problem :
  - Given an output $y = \mathbf{Q}\mathbf{x} + \mathbf{\Delta}$
  - Find the corresponding data set $\hat{\mathbf{x}} : \|\mathbf{Q}\hat{\mathbf{x}} - y\|_\infty \leq \delta_n$, and $d_{\text{Hamming}}(\mathbf{x}, \hat{\mathbf{x}}) \leq k_n$

The Equivalent Linear Inverse Problem

- Equivalent linear inverse problem :
  - Given an output $y = \mathbf{Q}\mathbf{x} + \boldsymbol{\Delta}$
  - Find the corresponding data set $\hat{\mathbf{x}} : \|\mathbf{Q}\hat{\mathbf{x}} - y\|_\infty \le \delta_n$ and $d_{\text{Hamming}}(\mathbf{x}, \hat{\mathbf{x}}) \le k_n$
- The noise level is δ_n, if the difference in each single query is at most δ_n

The Equivalent Linear Inverse Problem

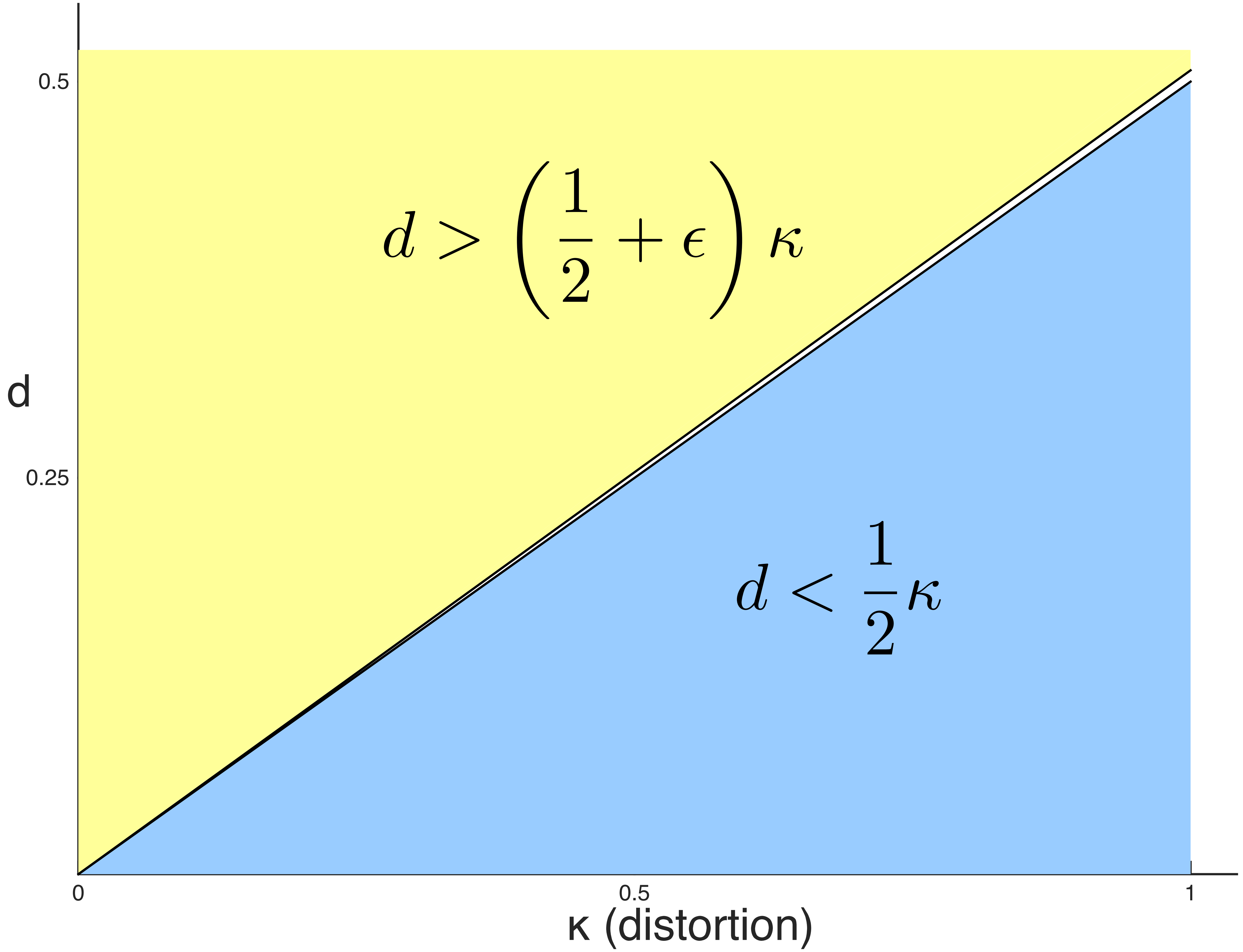
- Equivalent linear inverse problem :
  - Given an output $y = \mathbf{Q}\mathbf{x} + \mathbf{\Delta}$
  - Find the corresponding data set $\hat{\mathbf{x}} : \|\mathbf{Q}\hat{\mathbf{x}} - \boldsymbol{y}\|_\infty \leq \delta_n$ and $d_{\text{Hamming}}(\mathbf{x}, \hat{\mathbf{x}}) \leq k_n$
- The noise level is δ_n, if the difference in each single query is at most $\delta_n \Longleftrightarrow \|\mathbf{\Delta}\|_\infty \leq \delta_n (\Longleftrightarrow \forall i,\ \Delta_i \leq \delta_n)$

The Equivalent Linear Inverse Problem

- Equivalent linear inverse problem :
  - Given an output $y = \mathbf{Q}\mathbf{x} + \boldsymbol{\Delta}$
  - Find the corresponding data set $\hat{\mathbf{x}} : \|\mathbf{Q}\hat{\mathbf{x}} - \boldsymbol{y}\|_\infty \leq \delta_n$ and $d_{\text{Hamming}}(\mathbf{x}, \hat{\mathbf{x}}) \leq k_n$
- The noise level is δ_n, if the difference in each single query is at most $\delta_n \Longleftrightarrow \|\boldsymbol{\Delta}\|_\infty \leq \delta_n (\Longleftrightarrow \forall i,\ \Delta_i \leq \delta_n)$
- Query complexity $T_n^*(k_n, \delta_n)$: minimum number of queries required to extract data set within distortion k_n, under noise level δ_n

Main Result

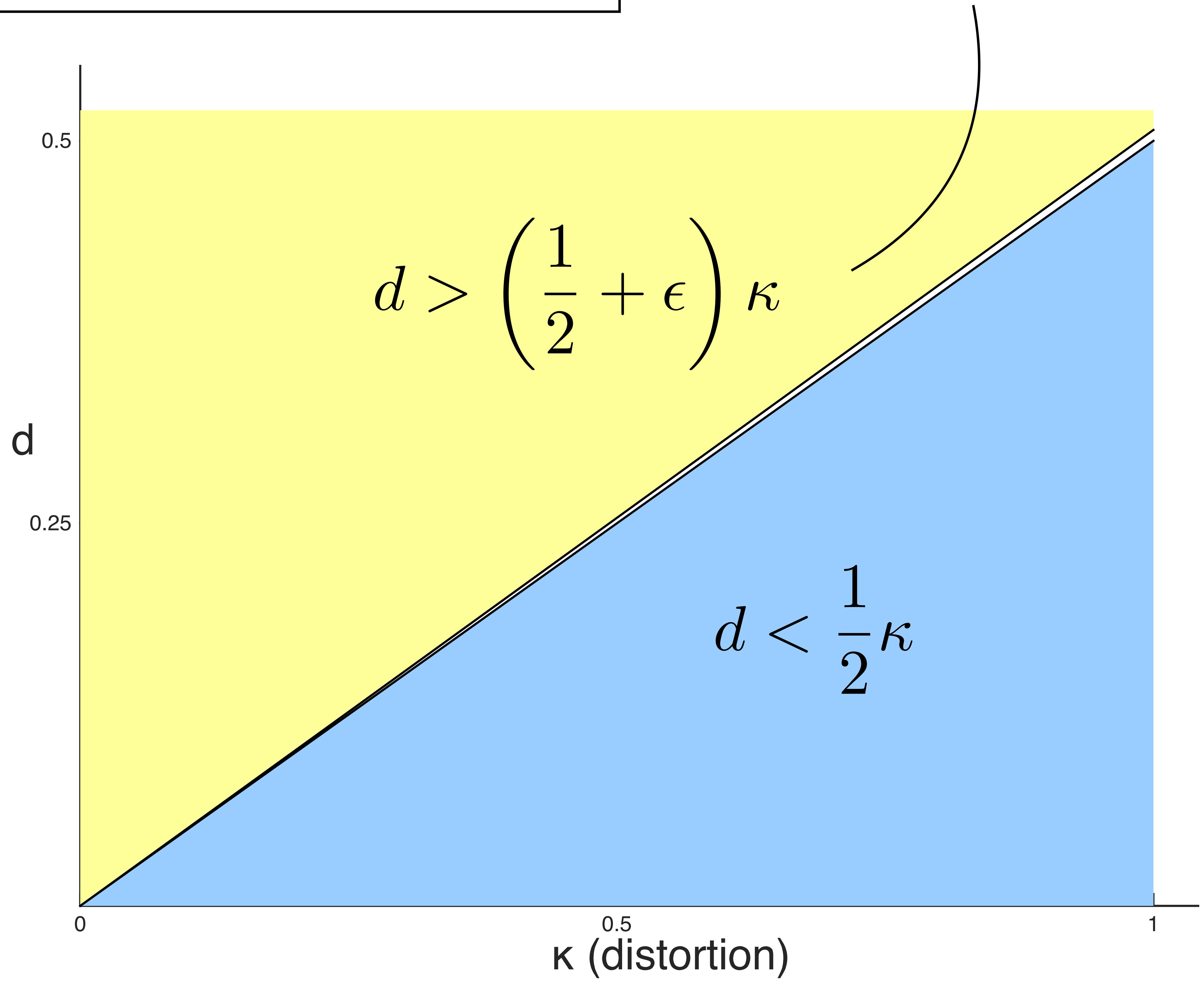
$$\delta_n = \Theta\left(n^d\right), \; k_n = \Theta\left(n^\kappa\right)$$



Main Result

$$\delta_n = \Theta\left(n^d\right), \ k_n = \Theta\left(n^\kappa\right)$$

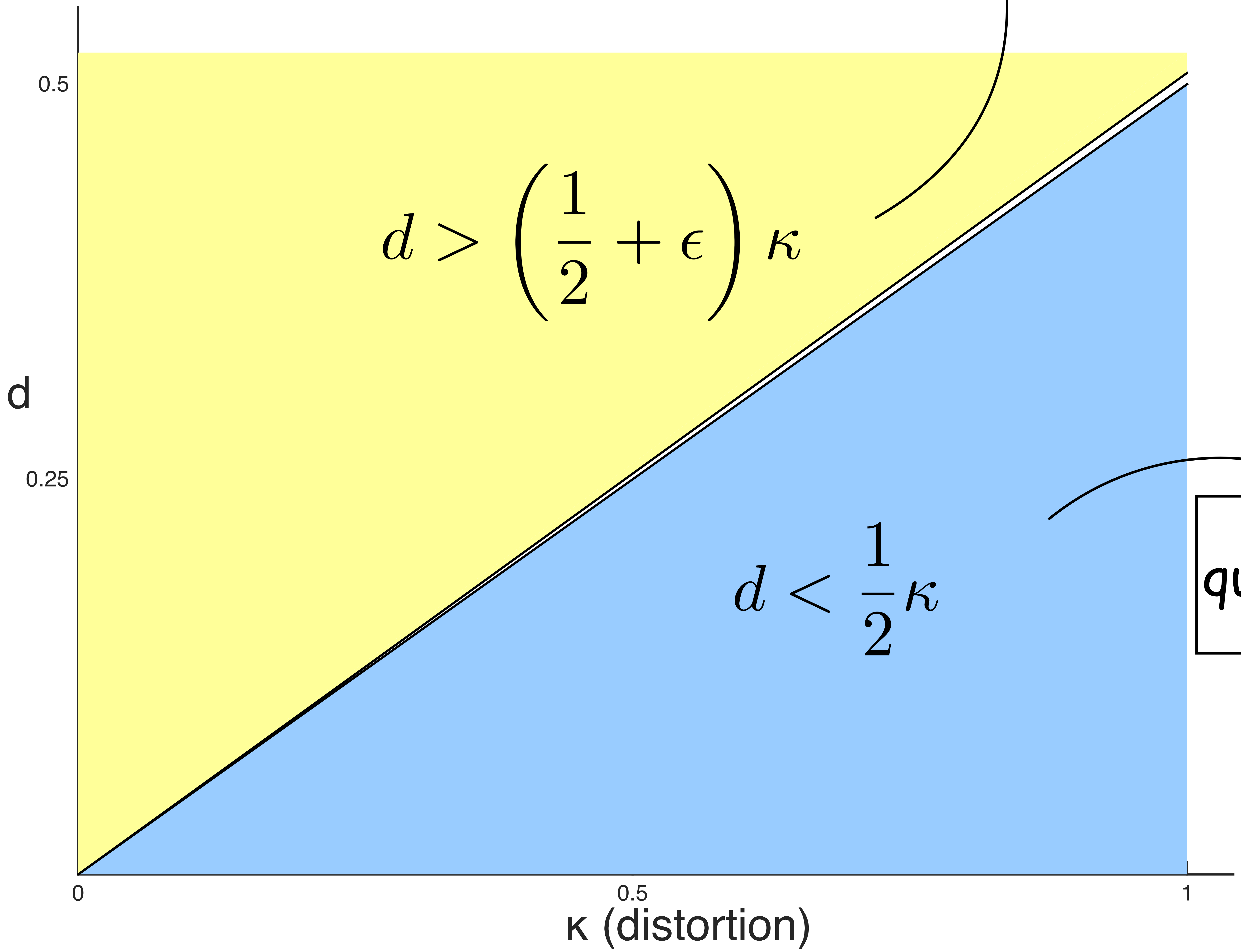
query complexity: non-polynomial $\Omega\left(\exp\left(n^\epsilon\right)\right)$



Main Result

$$\delta_n = \Theta\left(n^d\right), \; k_n = \Theta\left(n^\kappa\right)$$

query complexity: non-polynomial $\Omega\left(\exp\left(n^\epsilon\right)\right)$



$d > \left(\frac{1}{2} + \epsilon\right)\kappa$

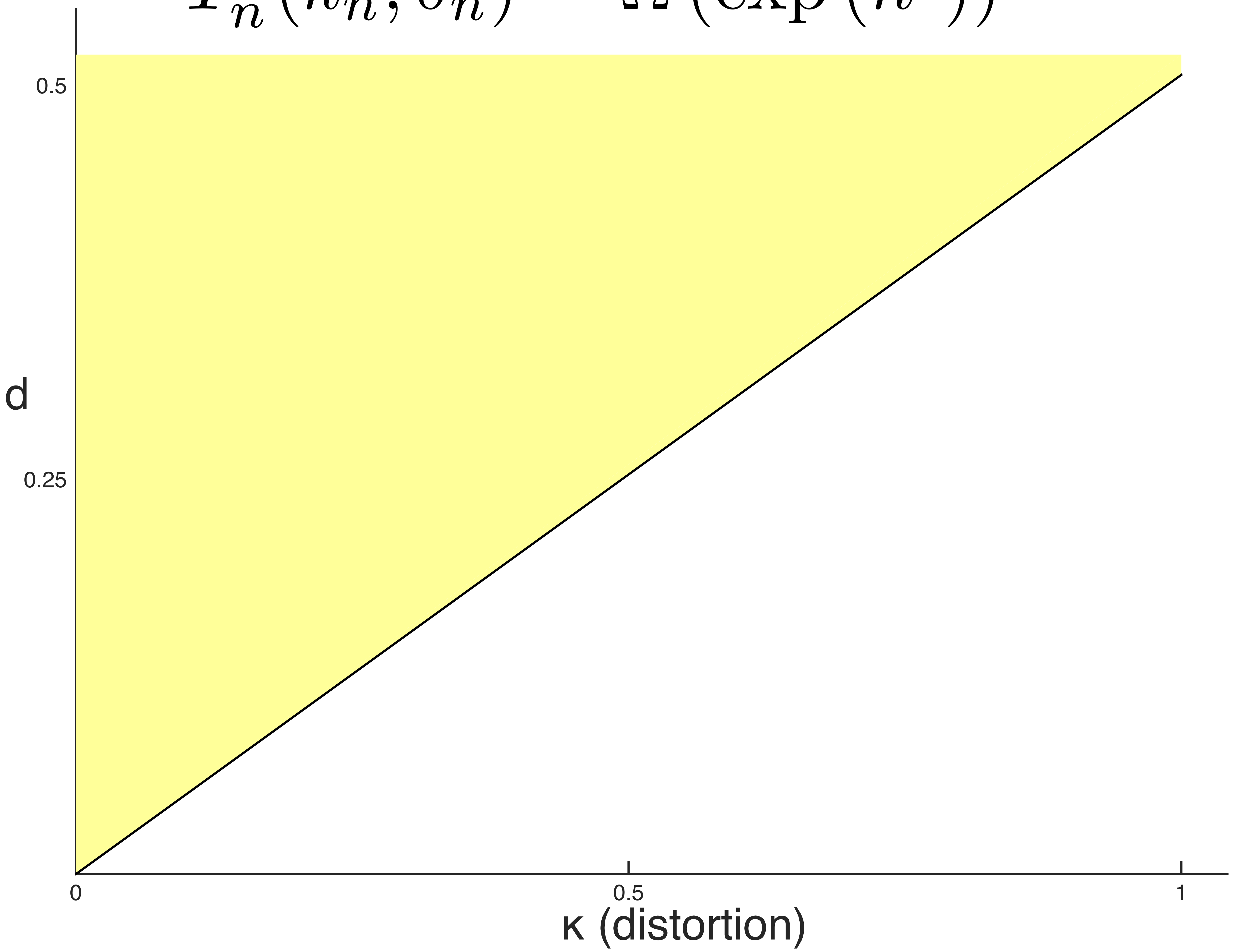
$d < \frac{1}{2}\kappa$

query complexity: sub-linear $\Theta\left(\frac{n}{\log n}\right)$

Regime 1: Impossibility of Polynomial Query Complexity

- **Regime 1:** $d > (\frac{1}{2} + \epsilon)\kappa, \text{ for any } \epsilon > 0$ (the noise is too large)

$$T_n^*(k_n, \delta_n) = \Omega\left(\exp\left(n^{\epsilon}\right)\right)$$



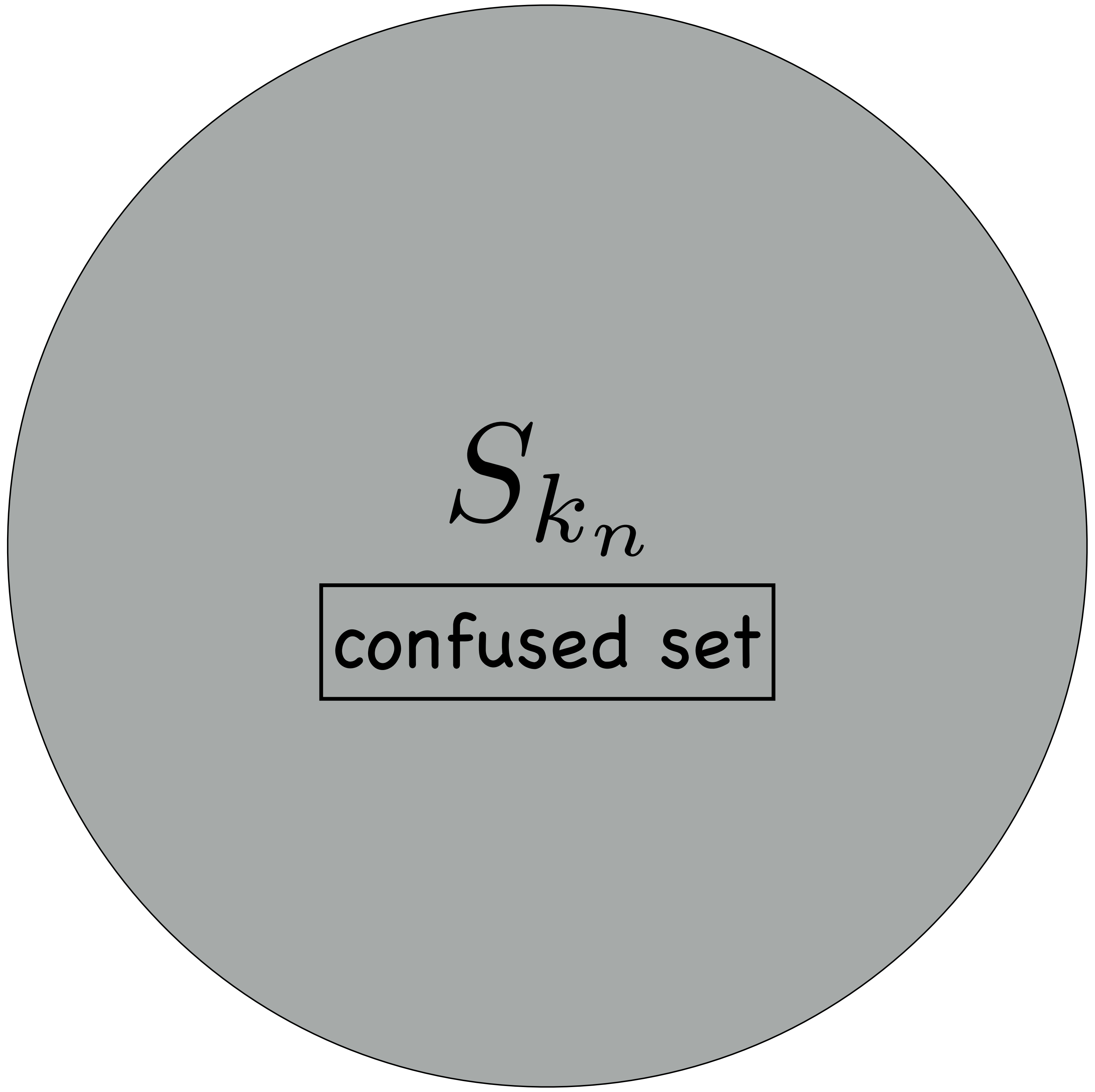
Impossibility of Polynomial Query

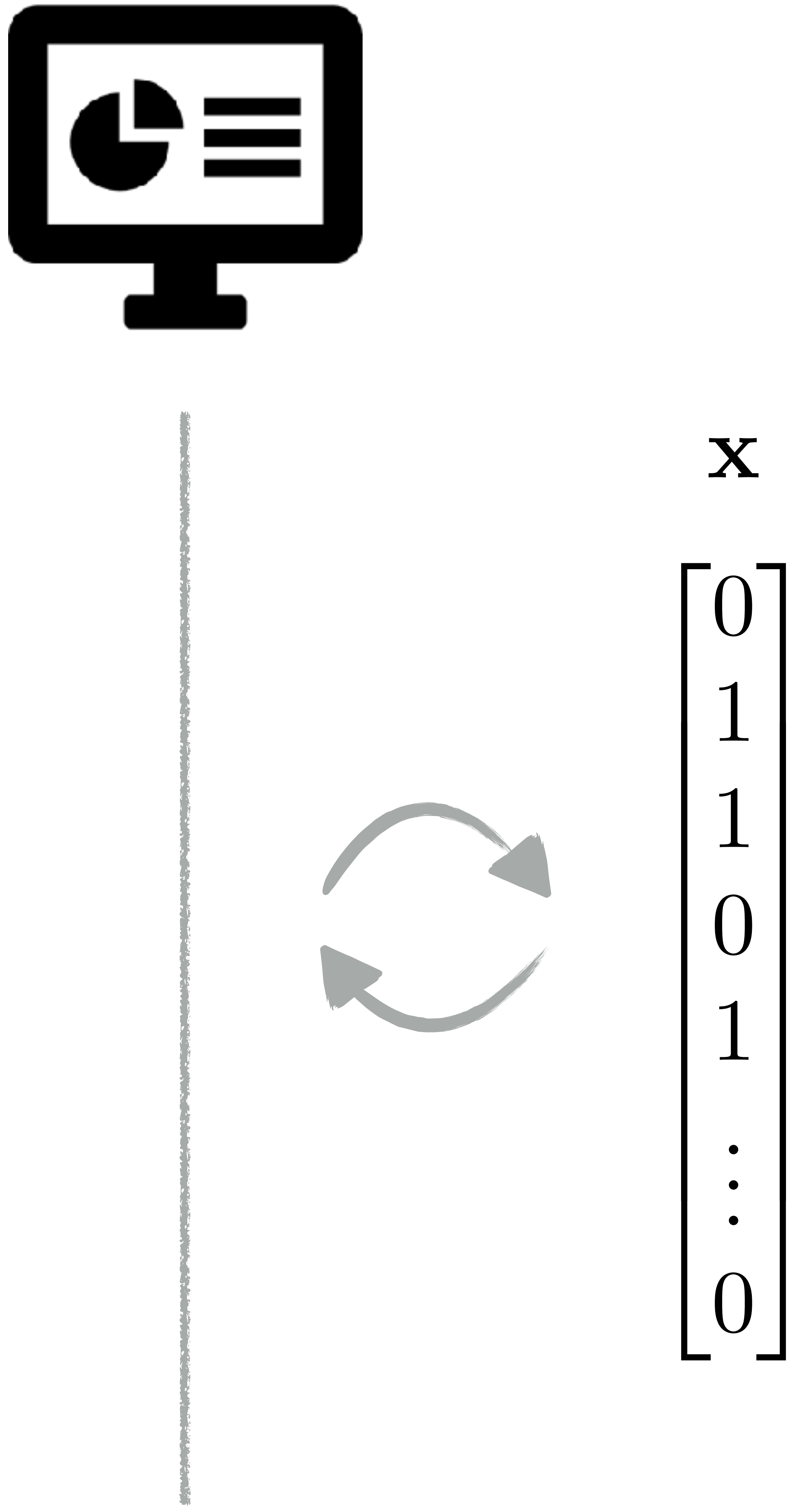
- Proof idea : without sufficient number of queries, there exists more than one possible data set which are consistent with the response.

Impossibility of Polynomial Query

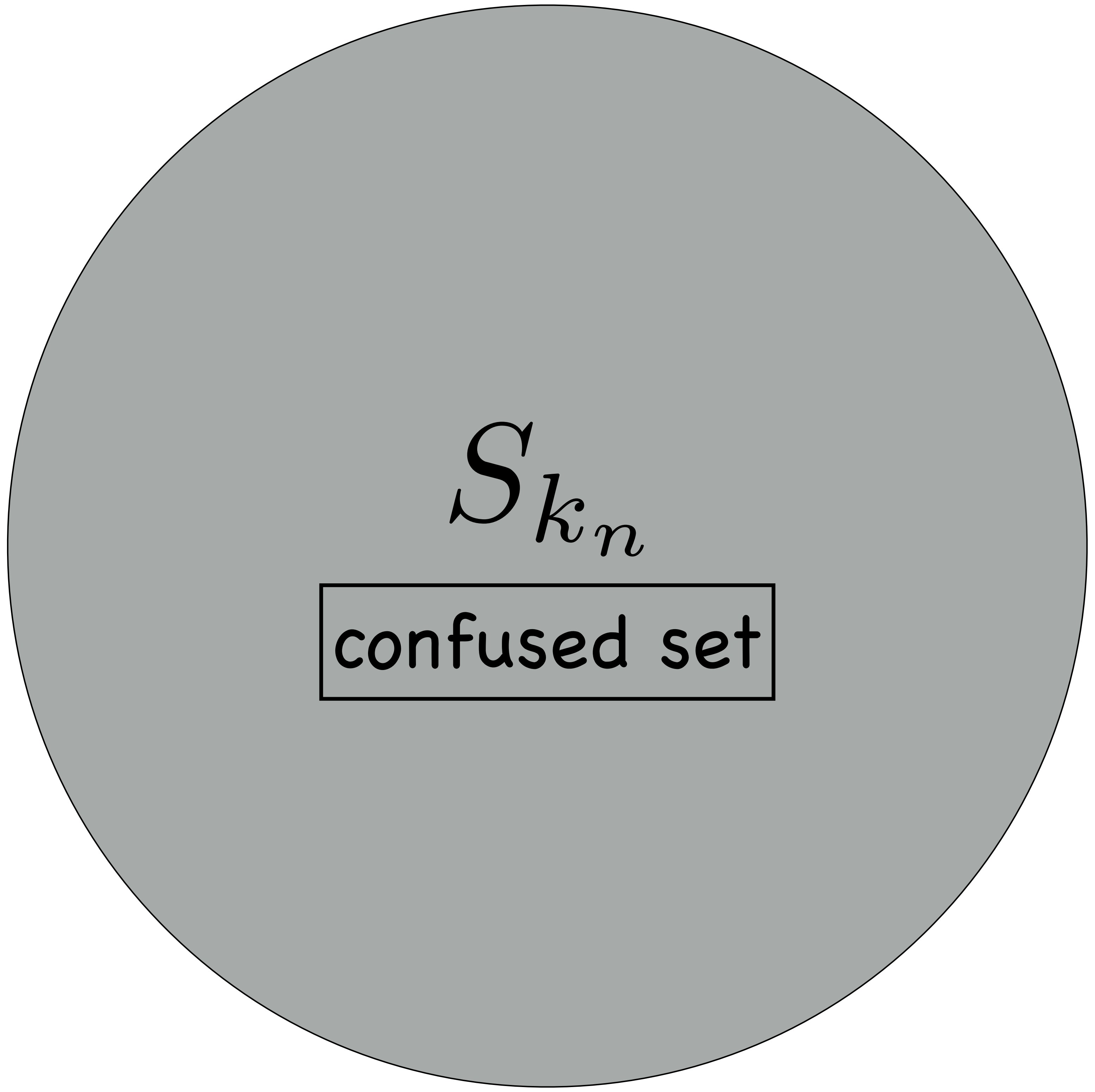
Impossibility of Polynomial Query

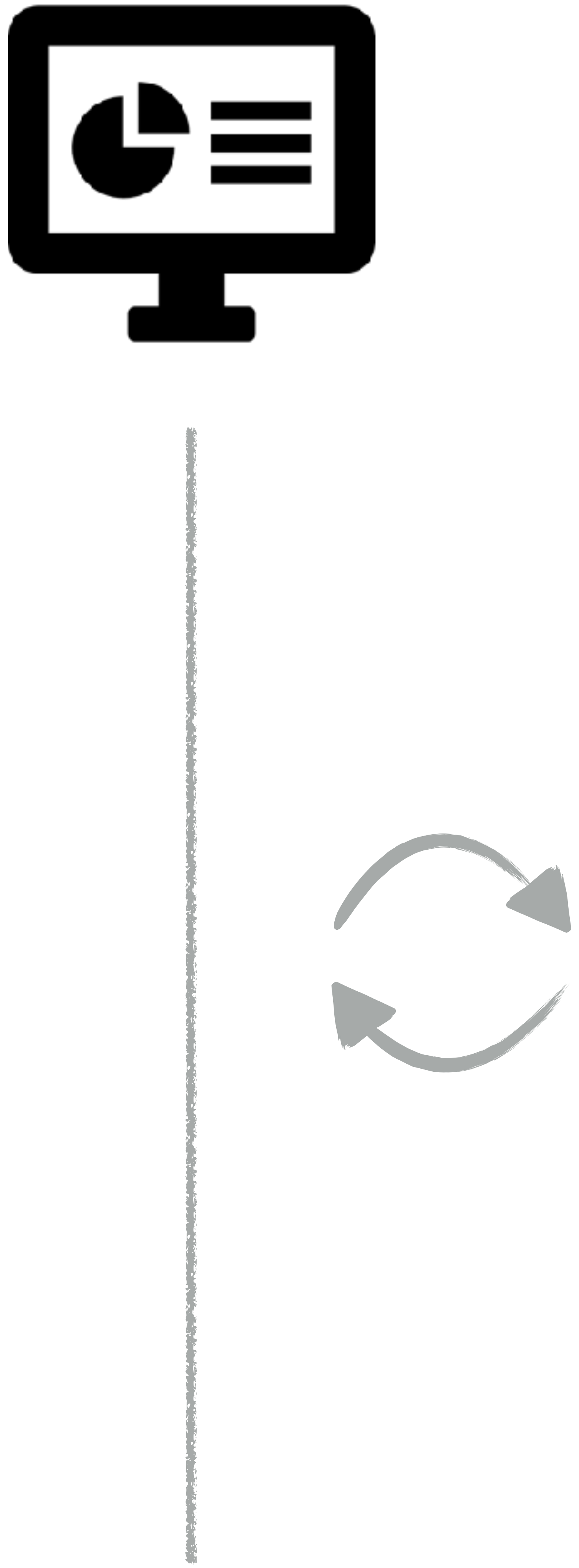
$$S_{k_n} \triangleq \{(\mathbf{x}, \tilde{\mathbf{x}}) \mid \mathbf{x}, \tilde{\mathbf{x}} \in \{0,1\}^n, \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 = k_n, \|\mathbf{x}\|_1 = \|\tilde{\mathbf{x}}\|_1\}$$





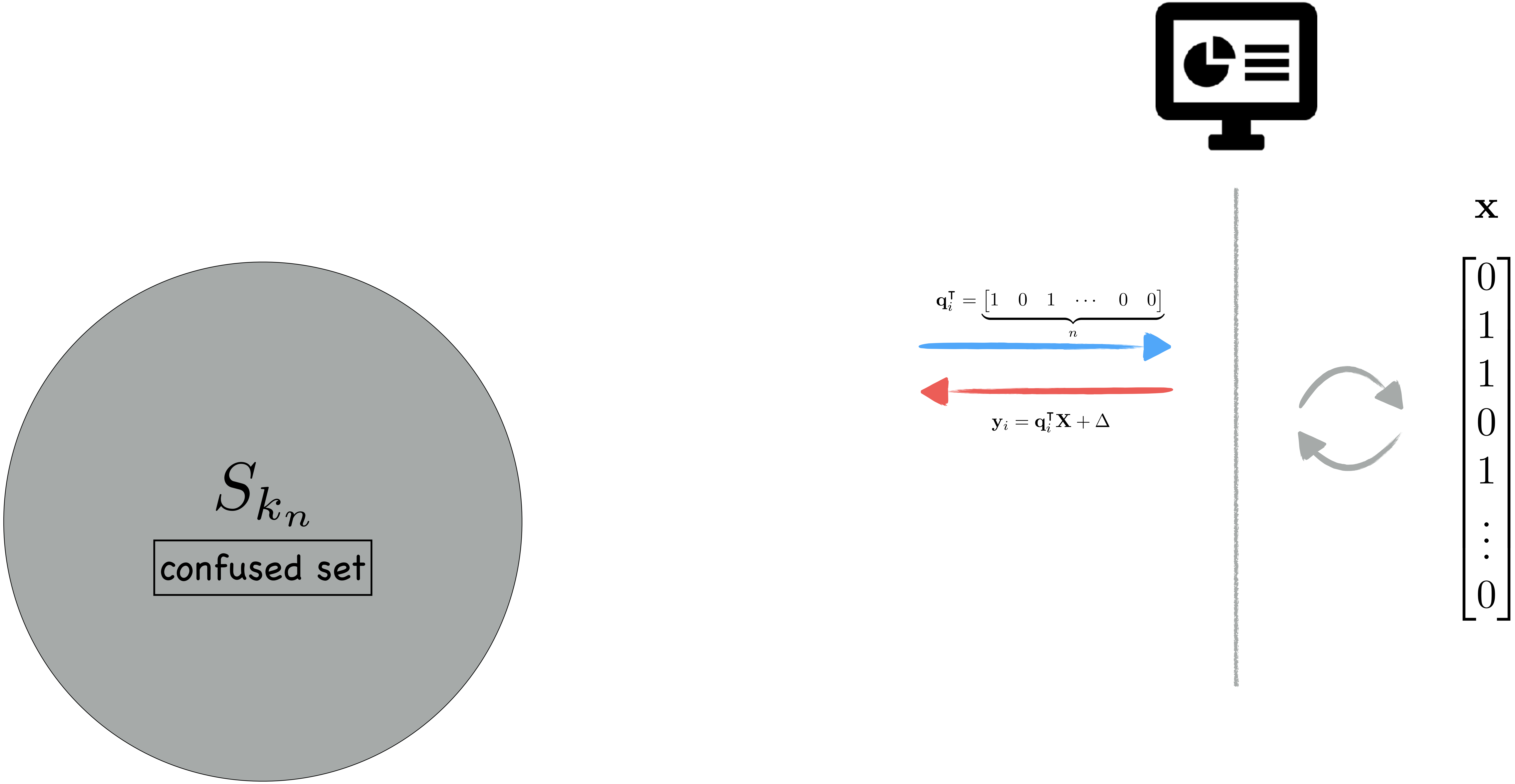
Impossibility of Polynomial Query



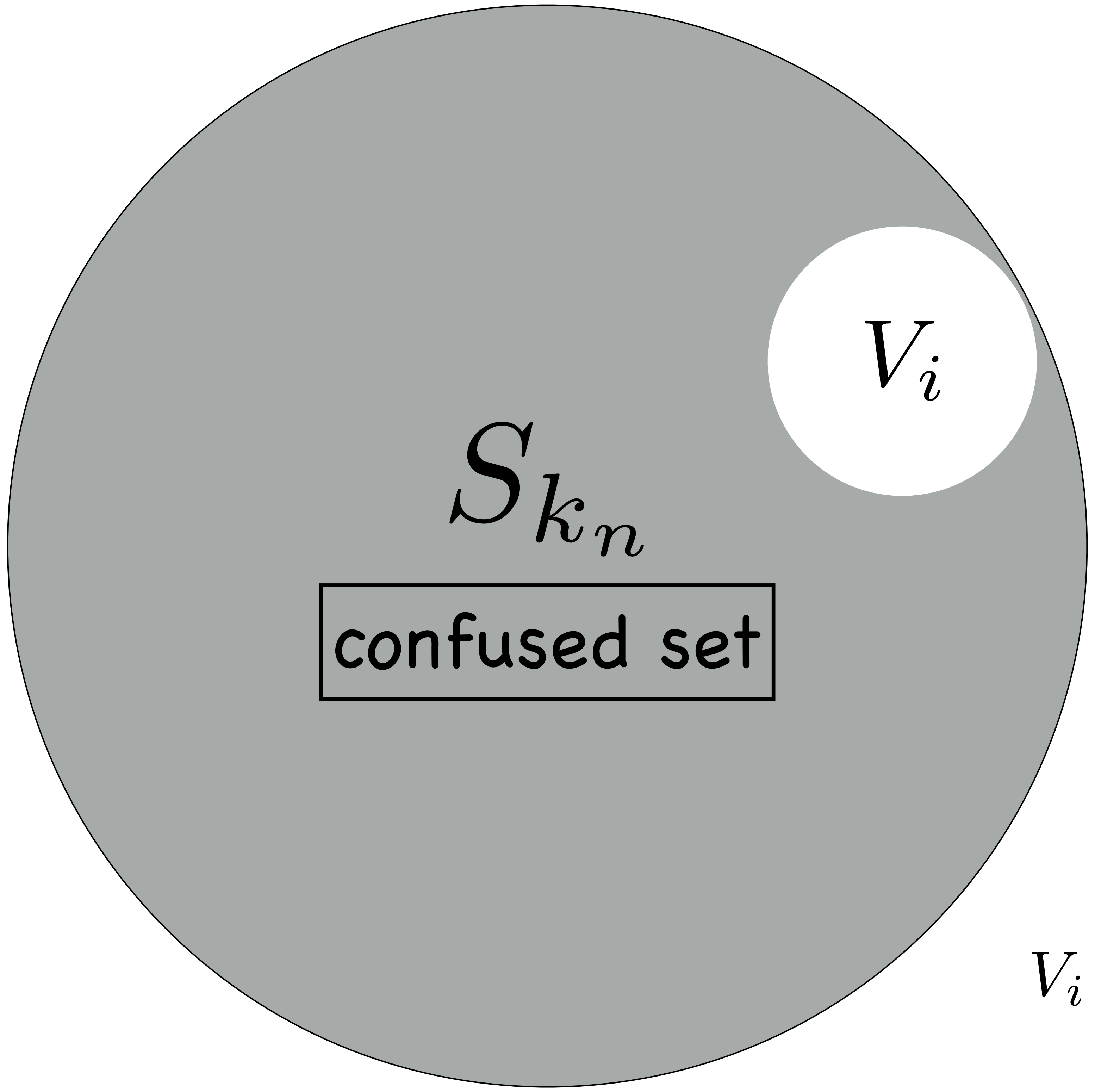


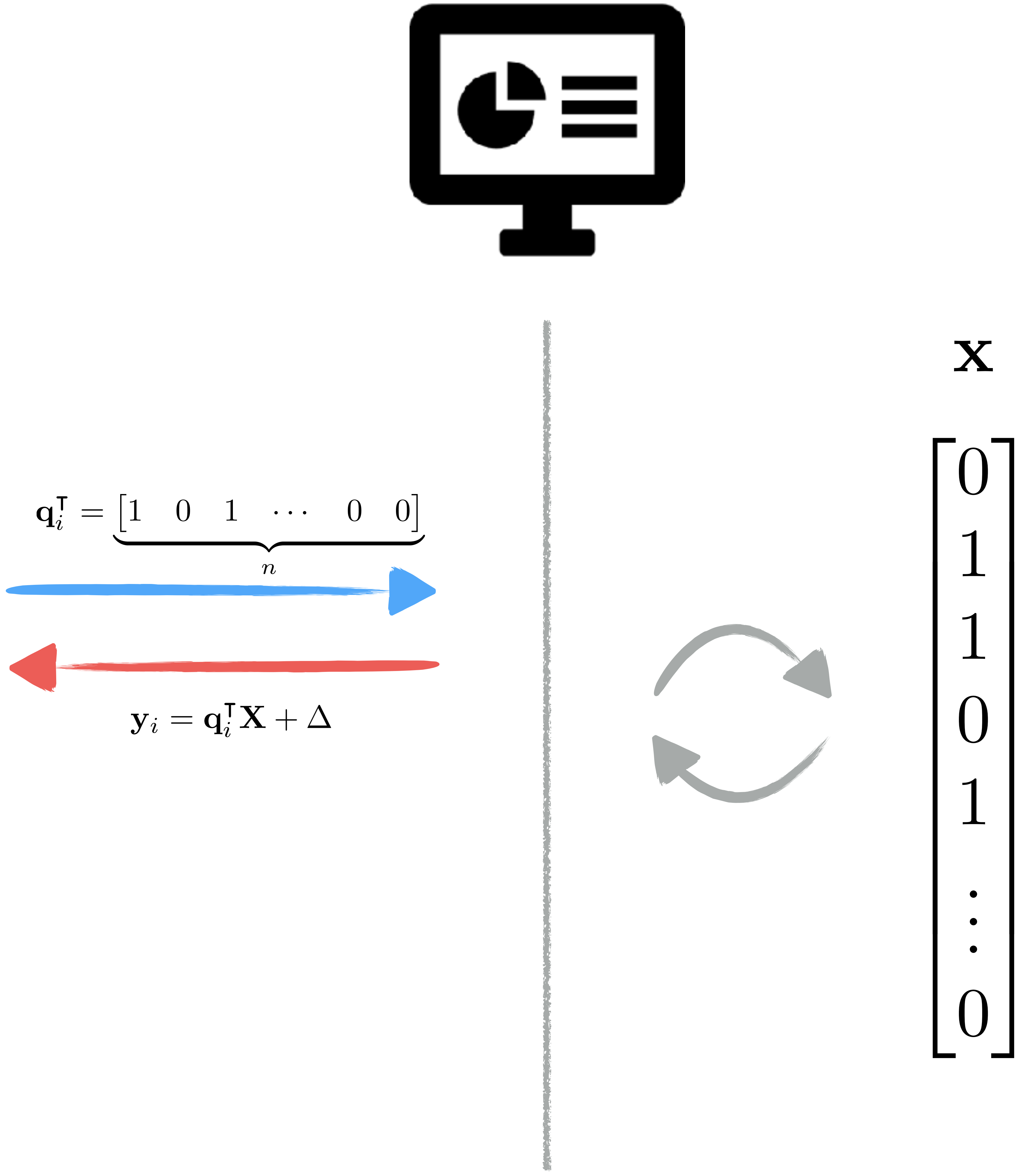
$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Impossibility of Polynomial Query



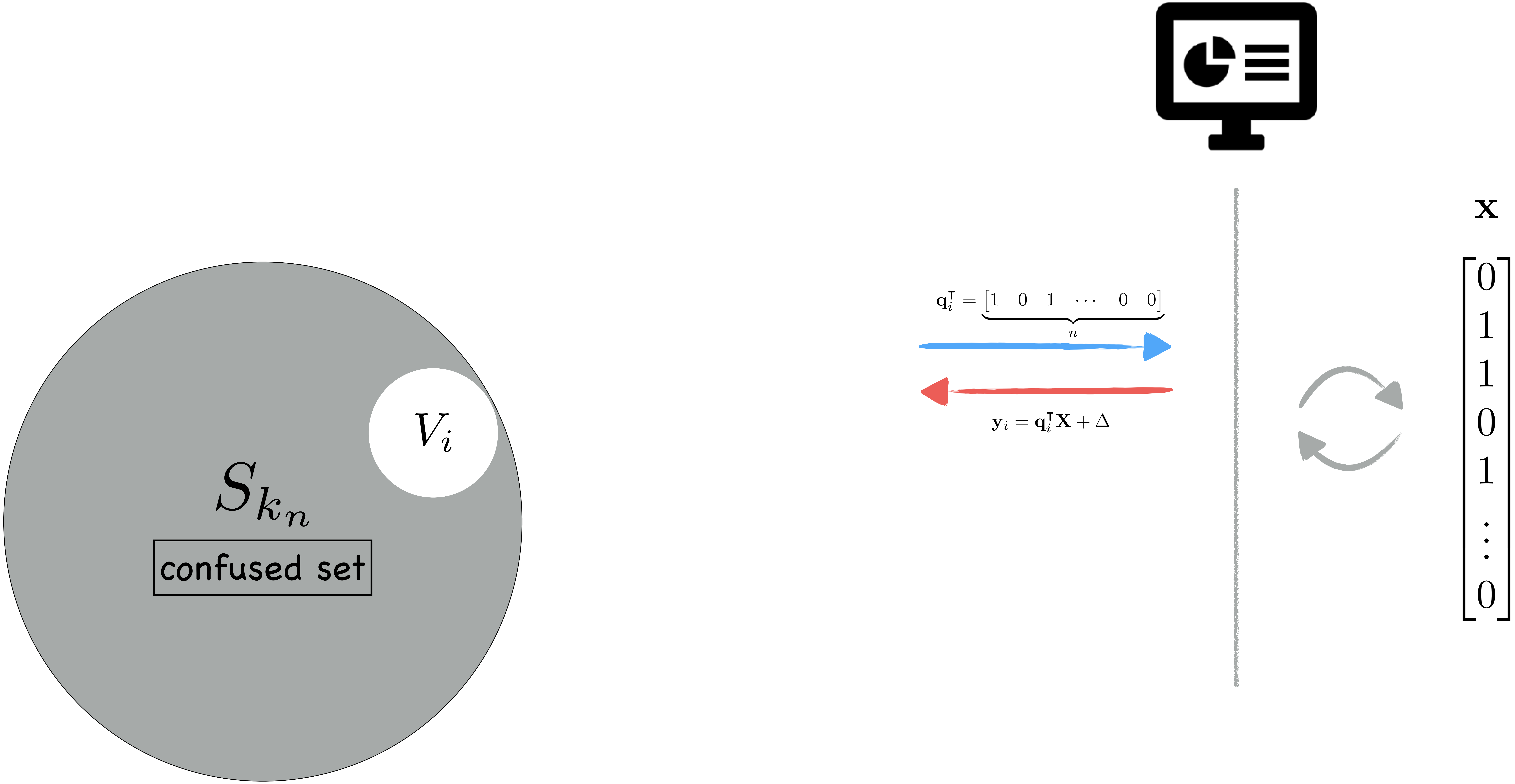
Impossibility of Polynomial Query



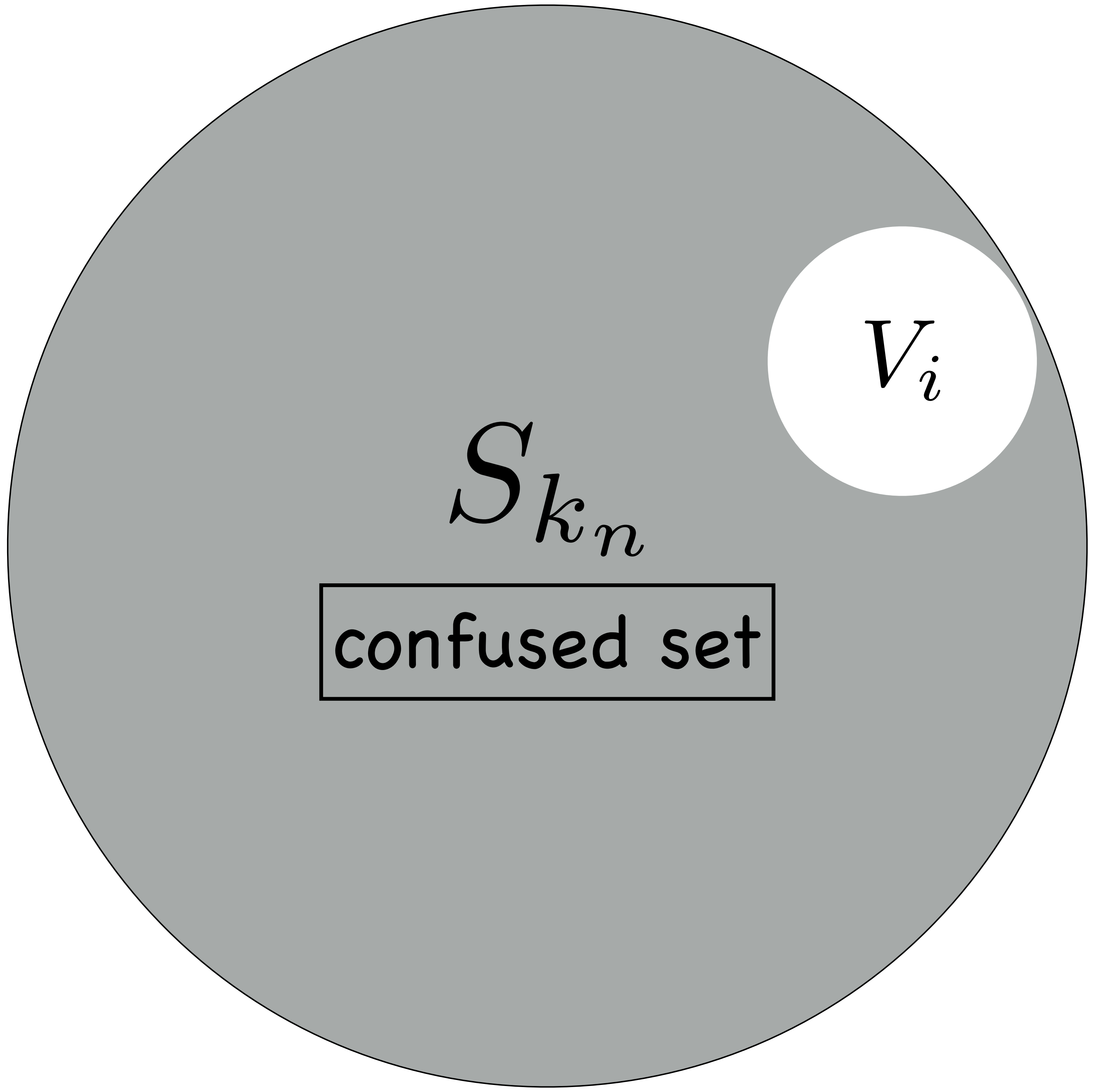


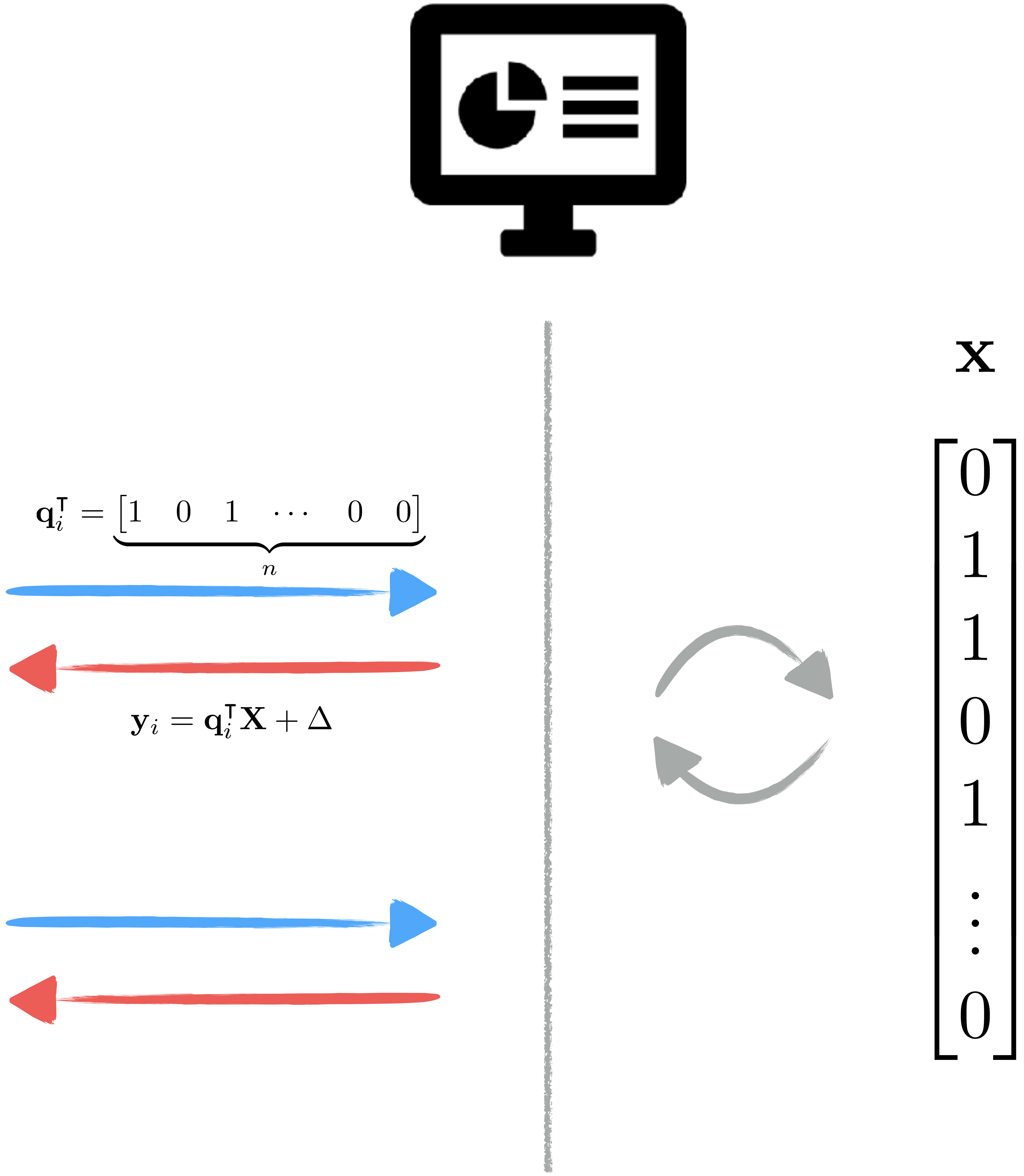
$$V_i \triangleq \{(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n} \mid |\mathbf{q}_i \cdot (\mathbf{x} - \tilde{\mathbf{x}})| > \delta_n\}.$$

Impossibility of Polynomial Query

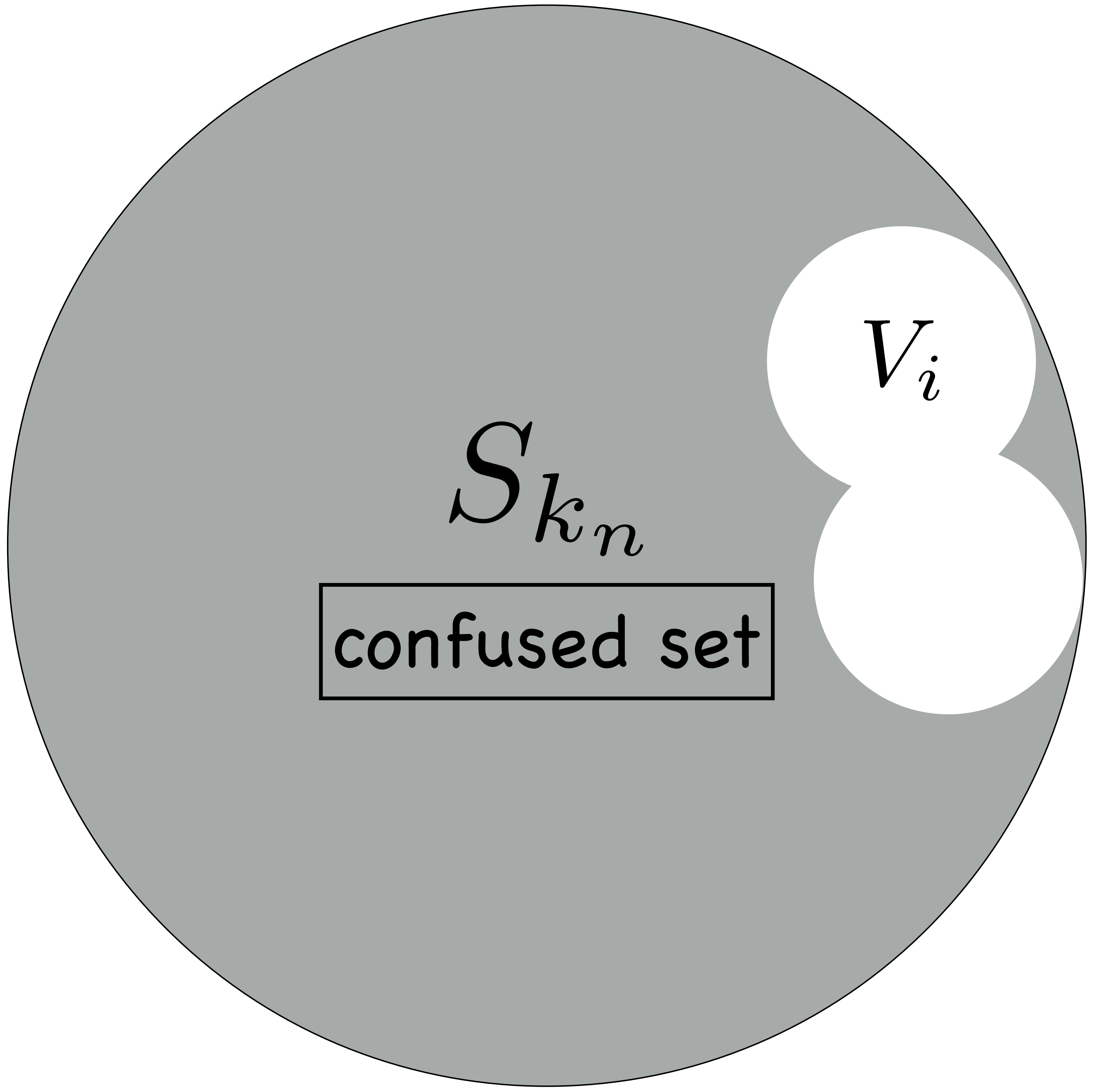


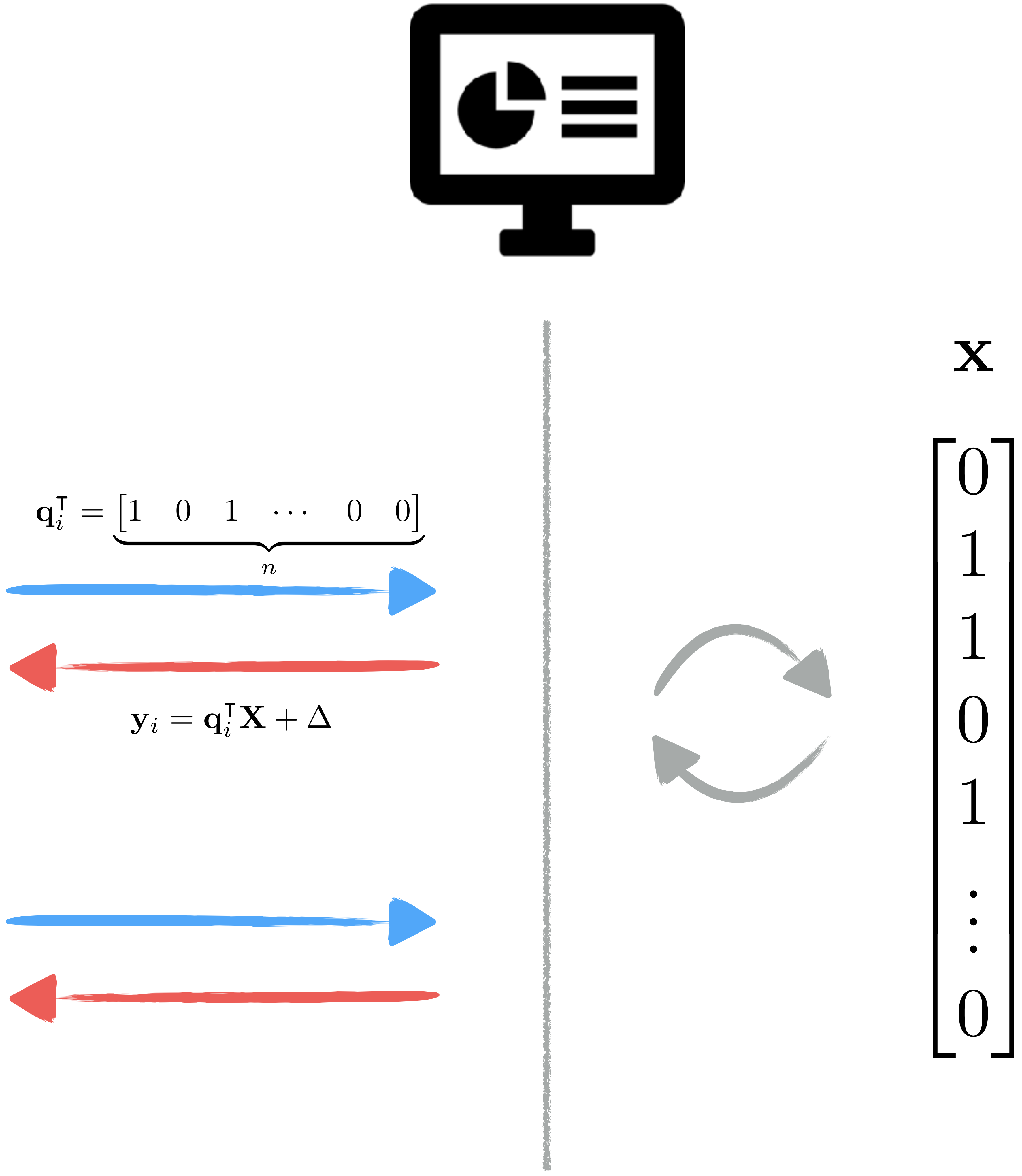
Impossibility of Polynomial Query



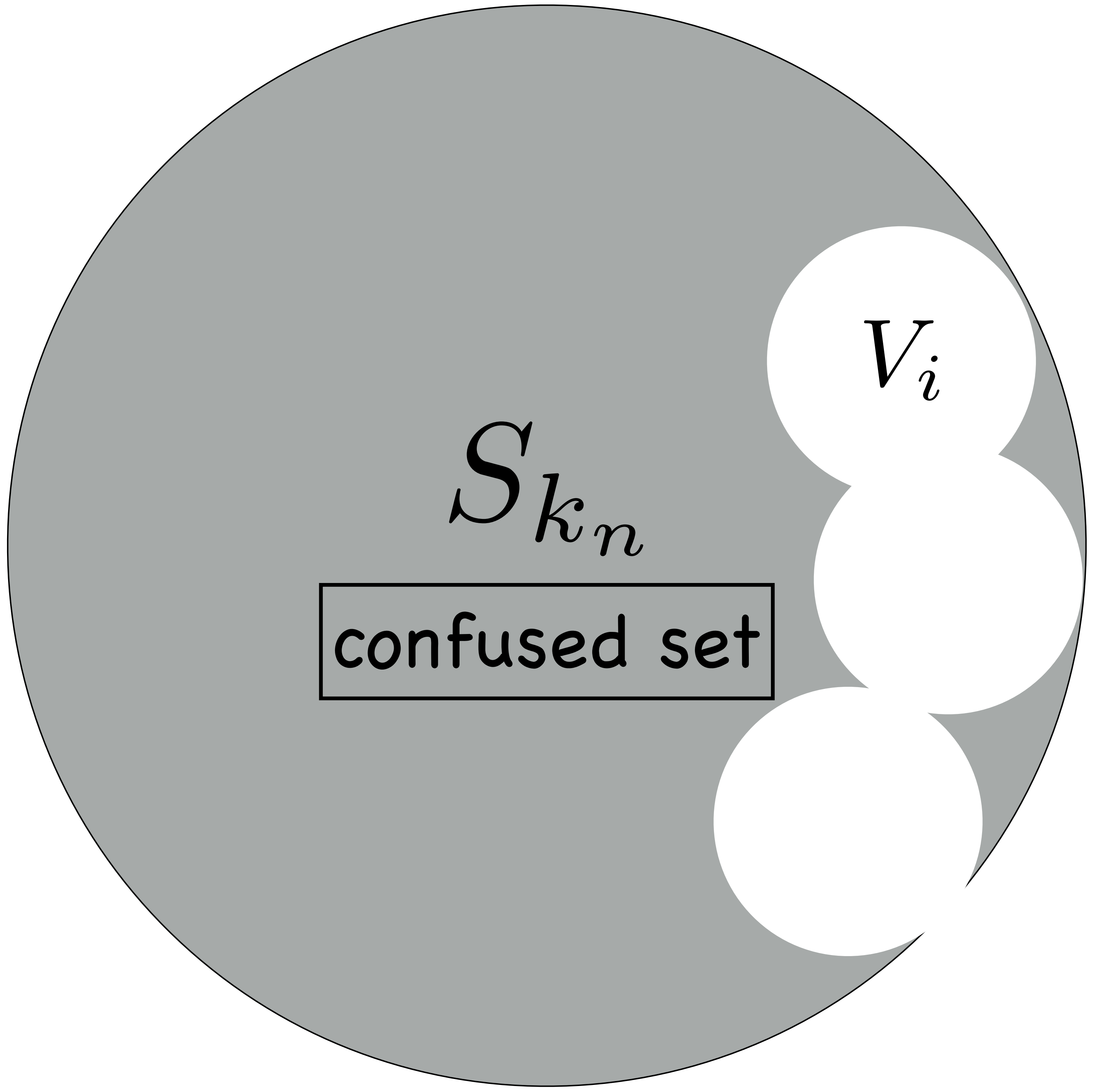


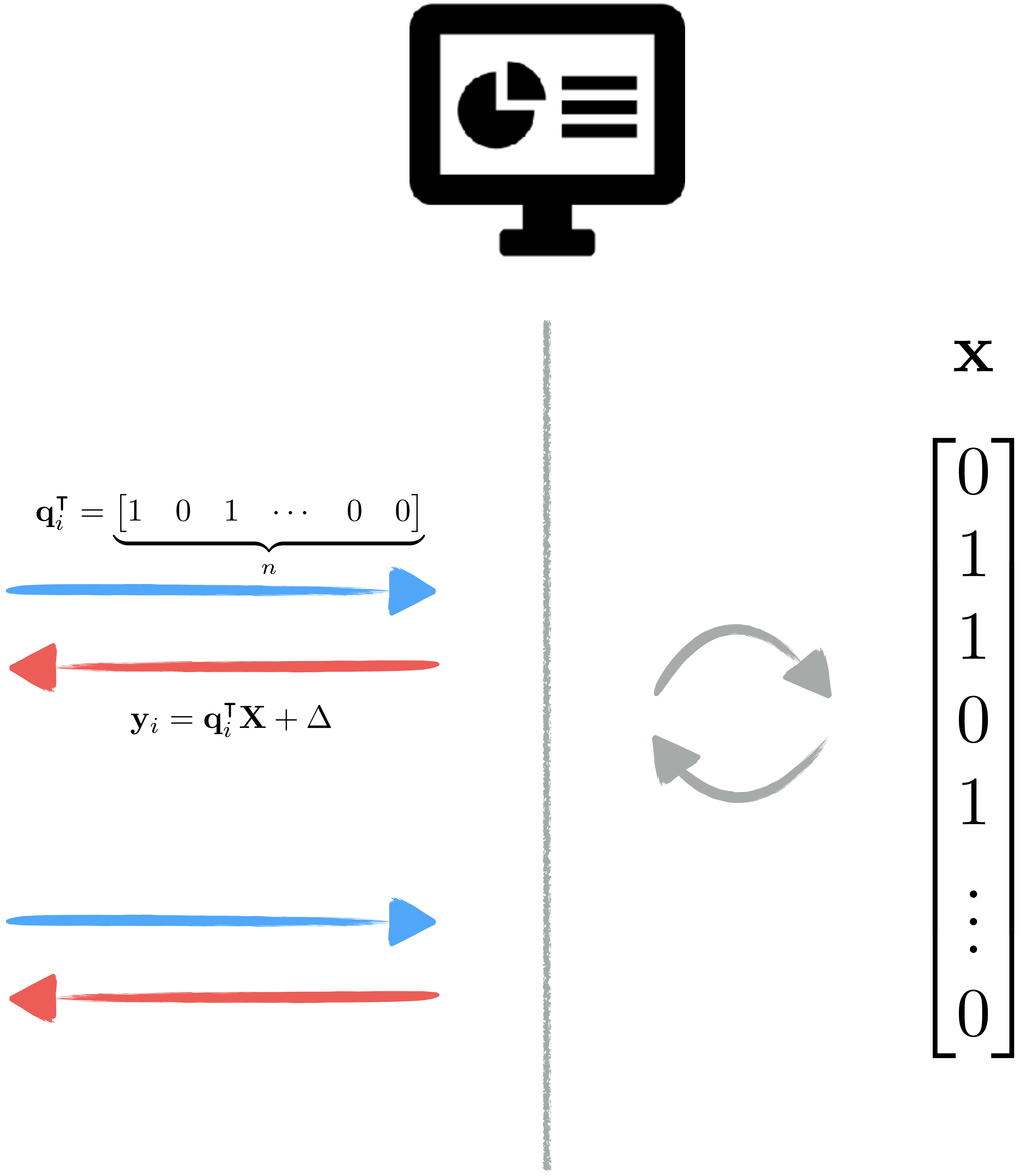
Impossibility of Polynomial Query



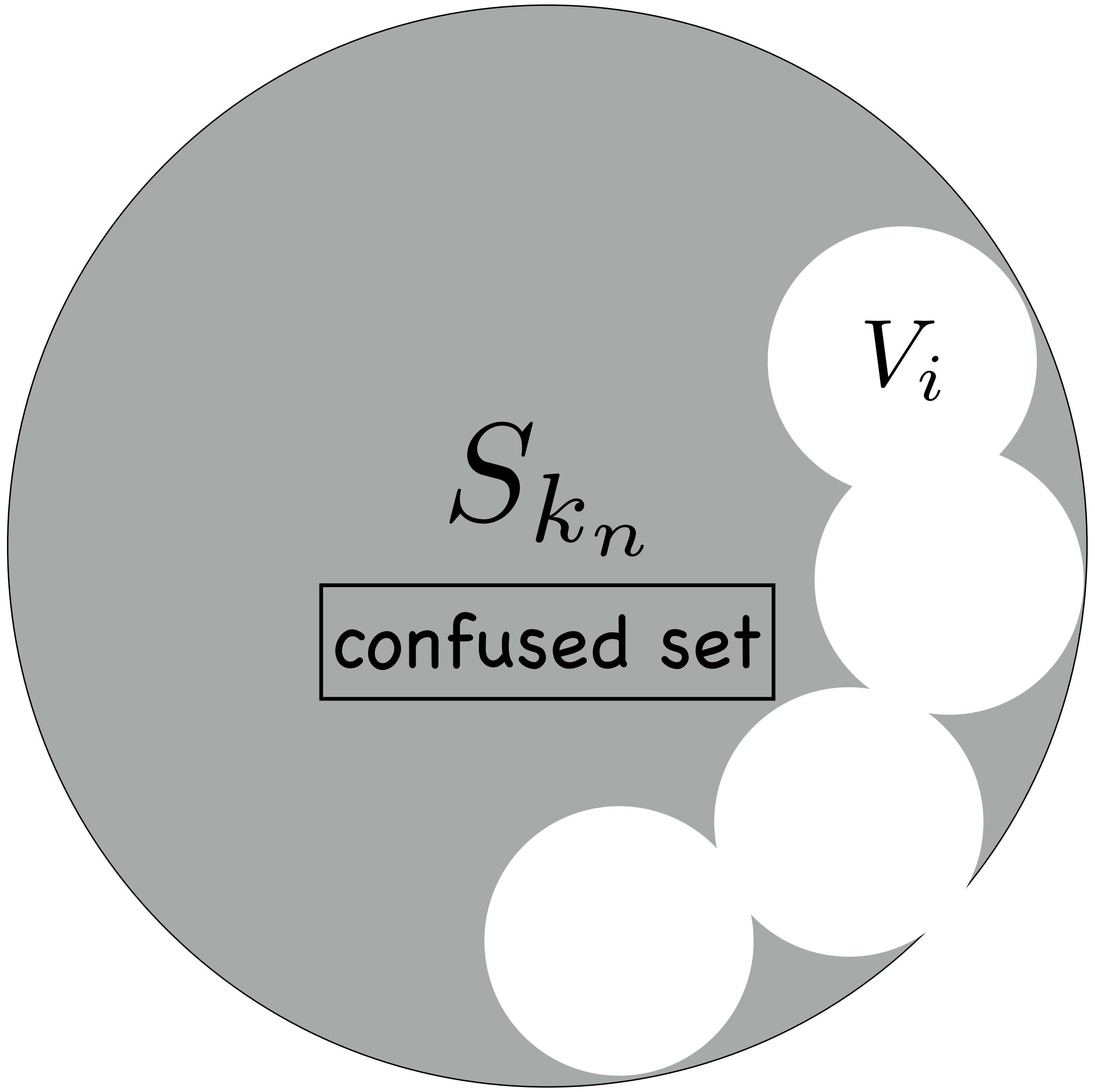


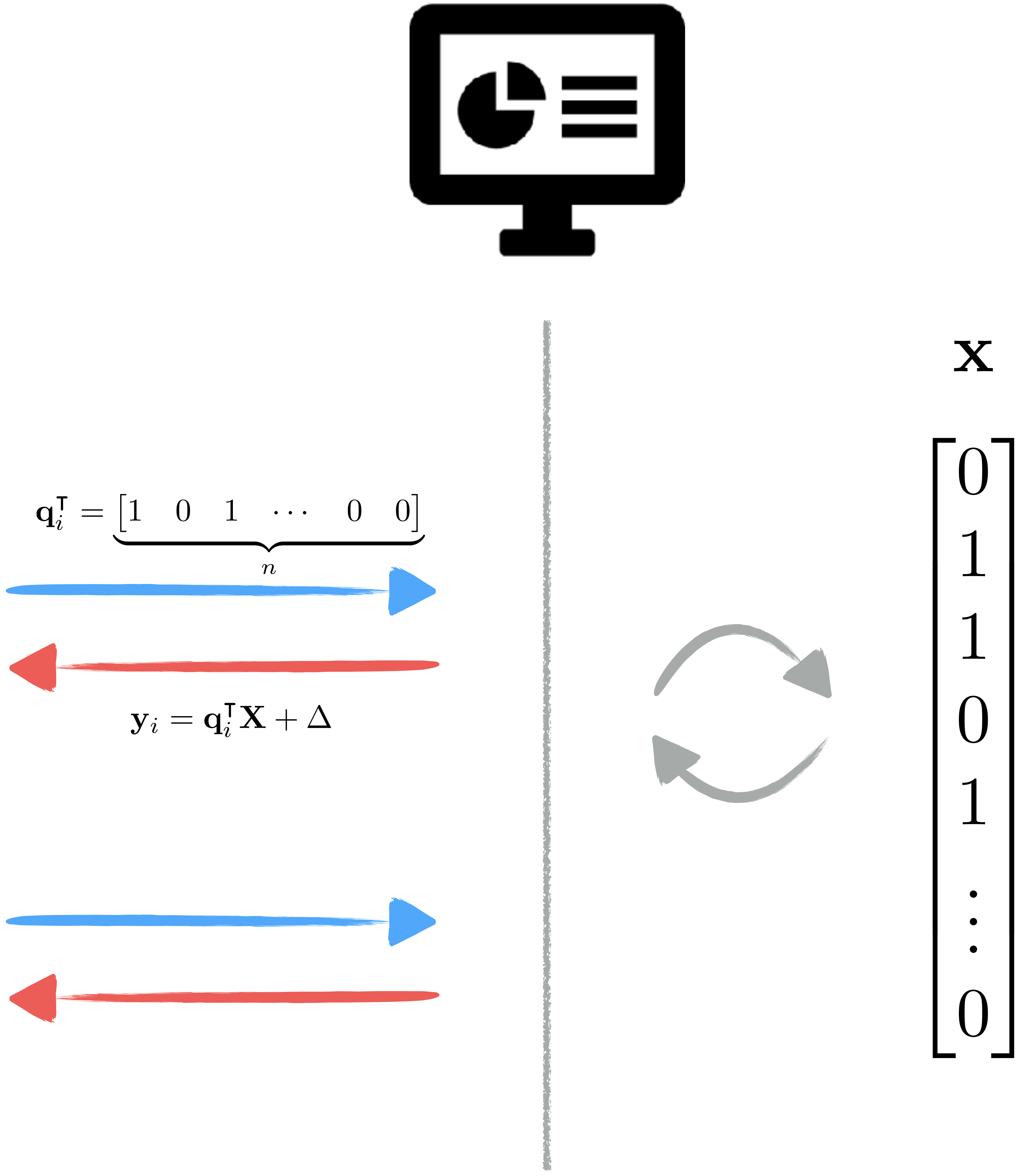
Impossibility of Polynomial Query



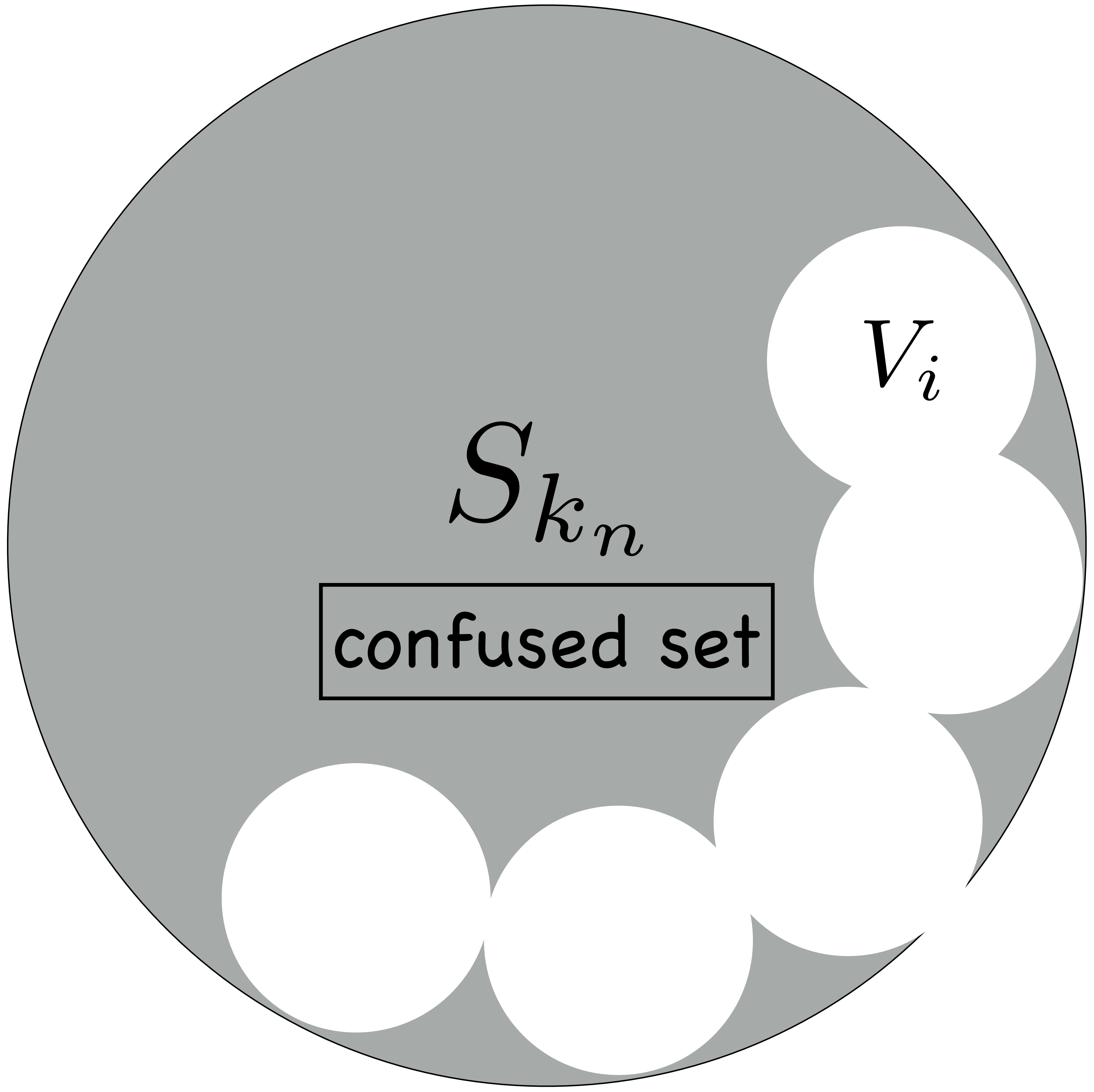


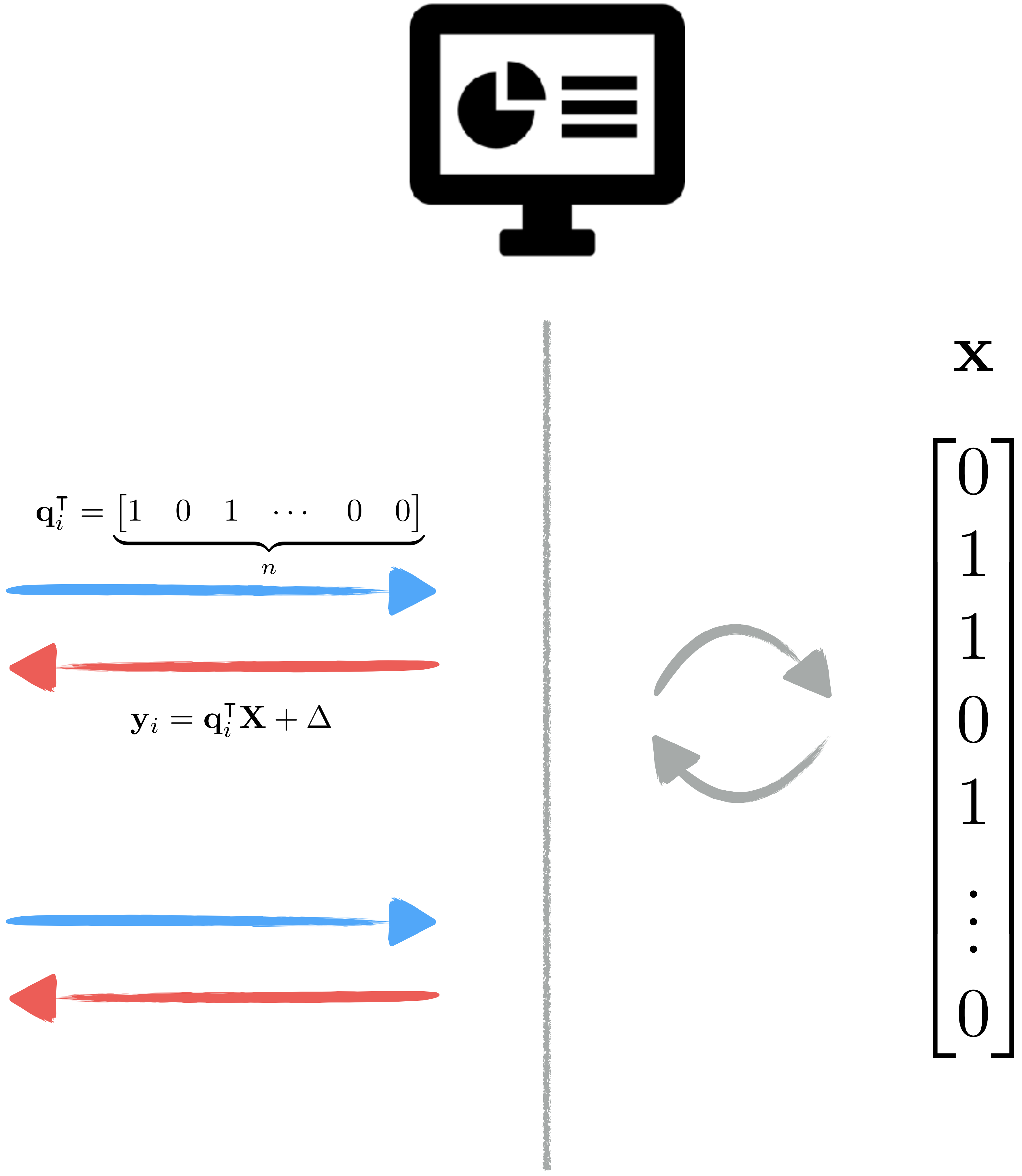
Impossibility of Polynomial Query



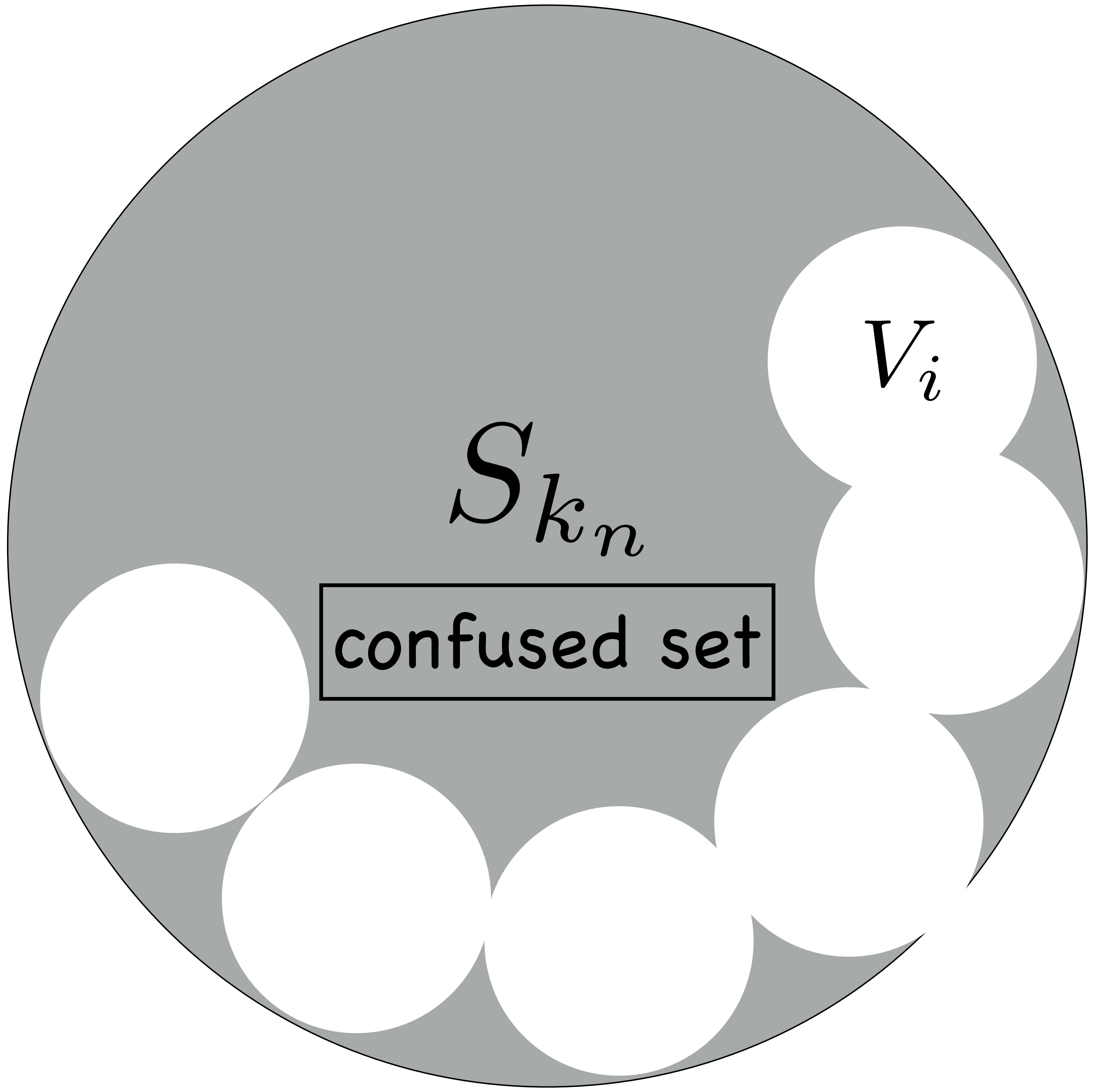


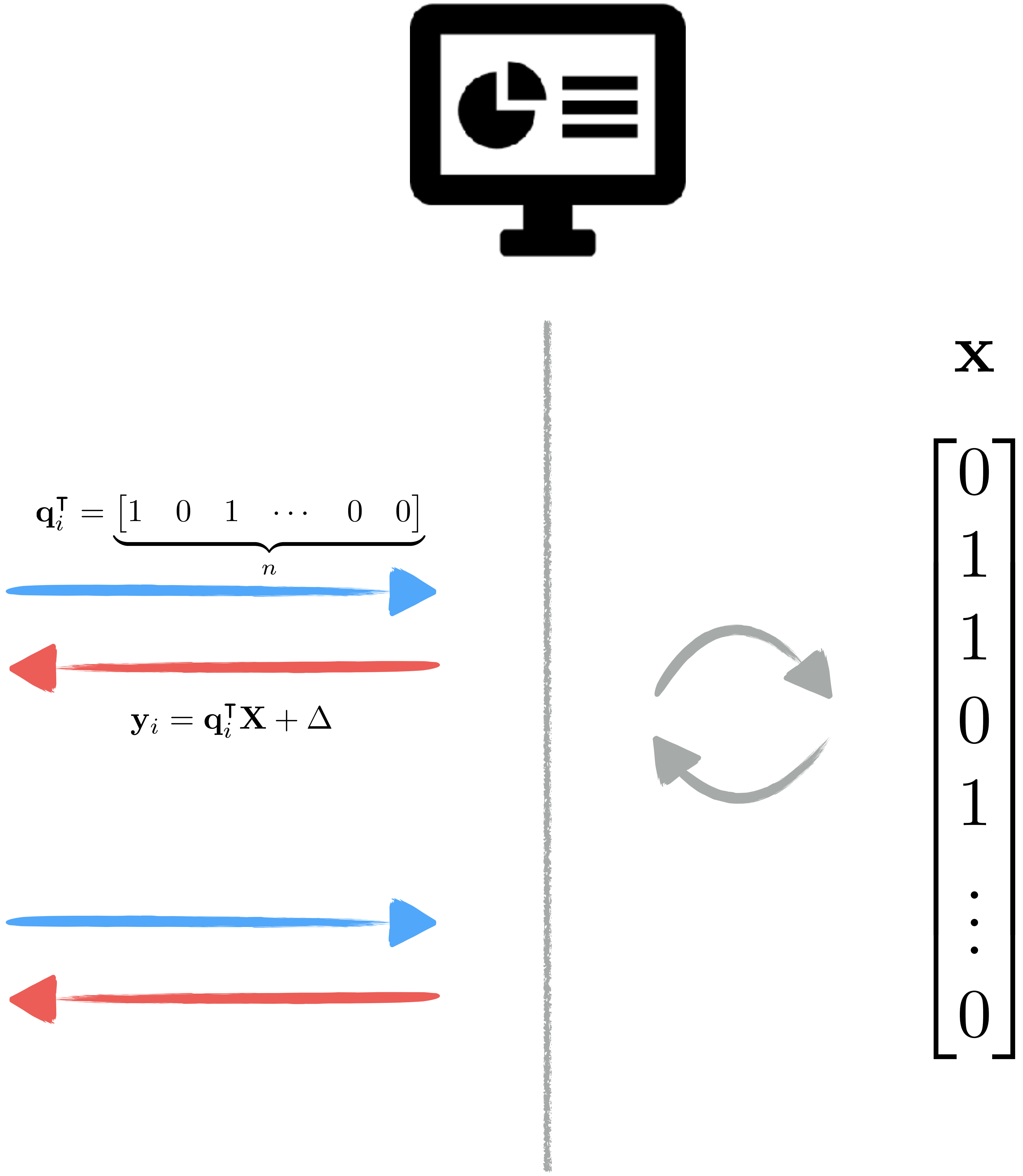
Impossibility of Polynomial Query



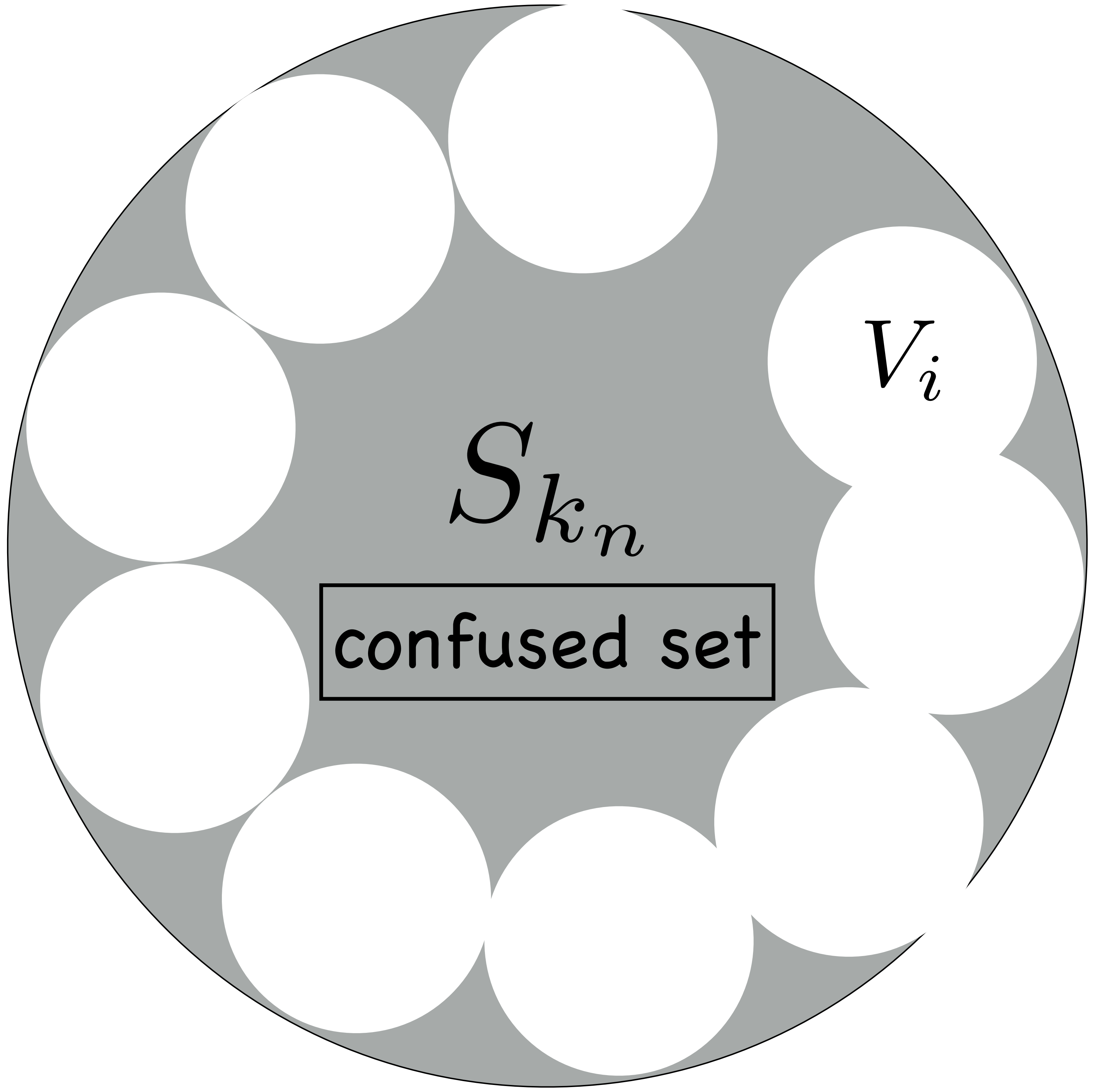


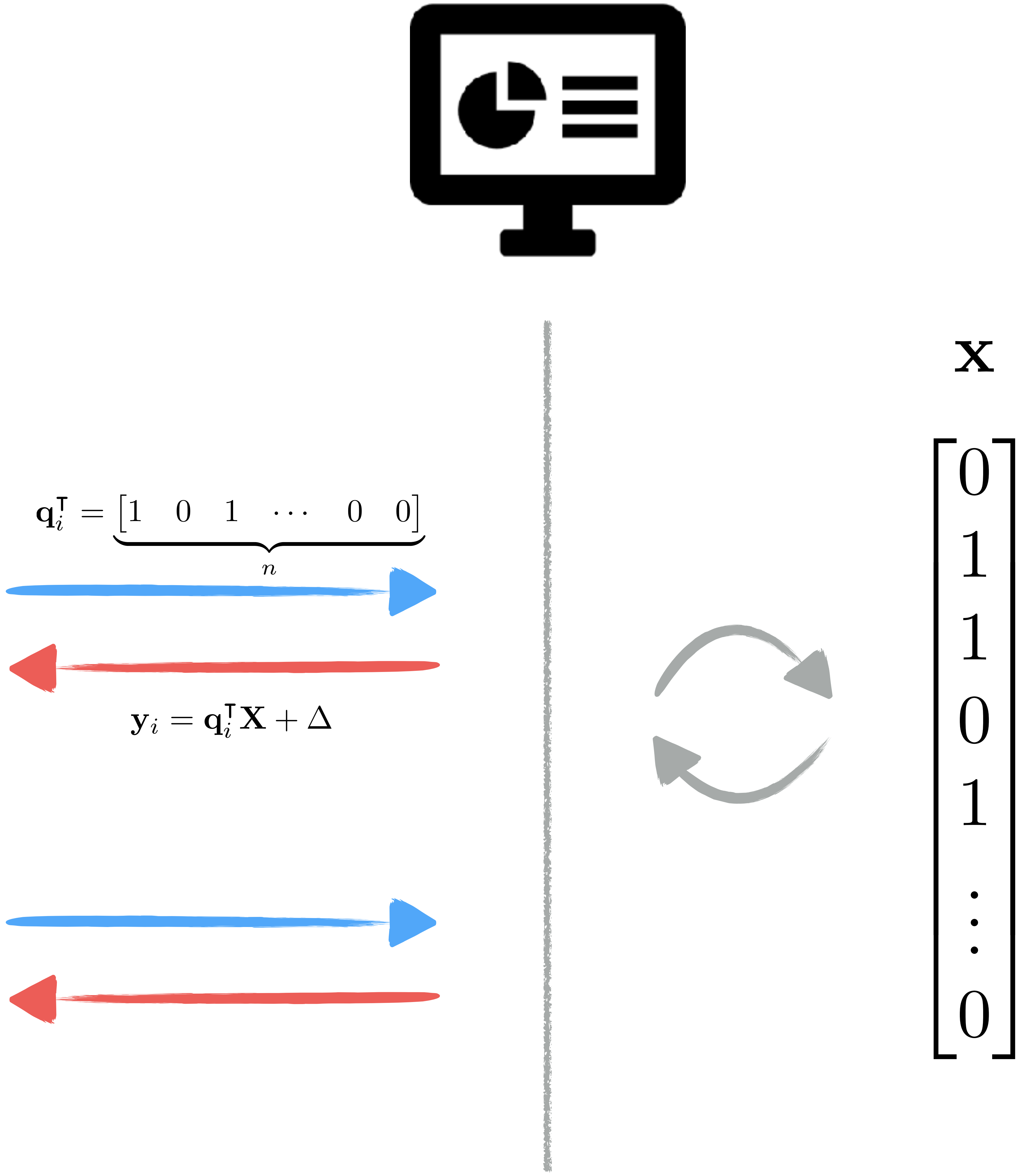
Impossibility of Polynomial Query





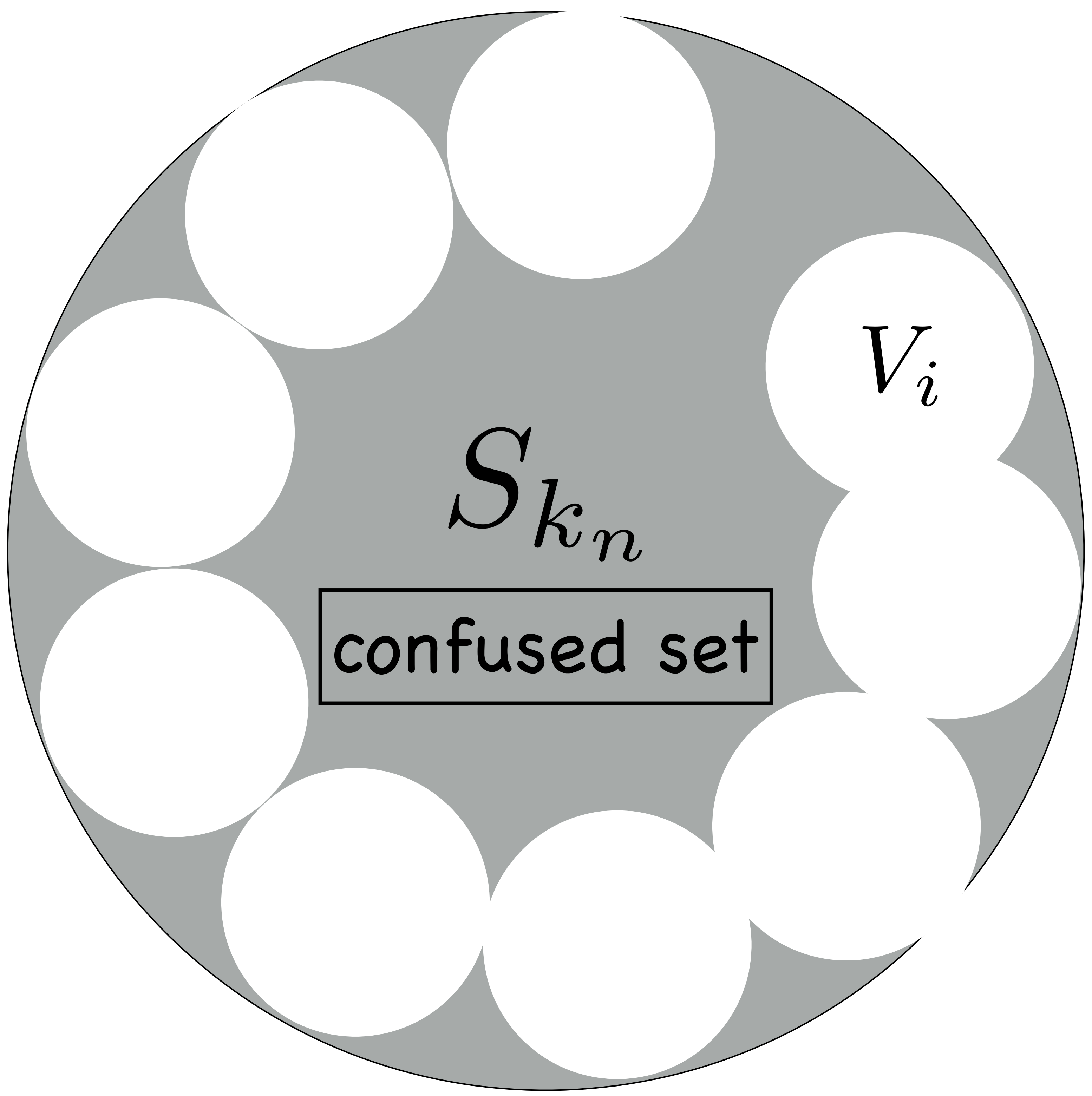
Impossibility of Polynomial Query

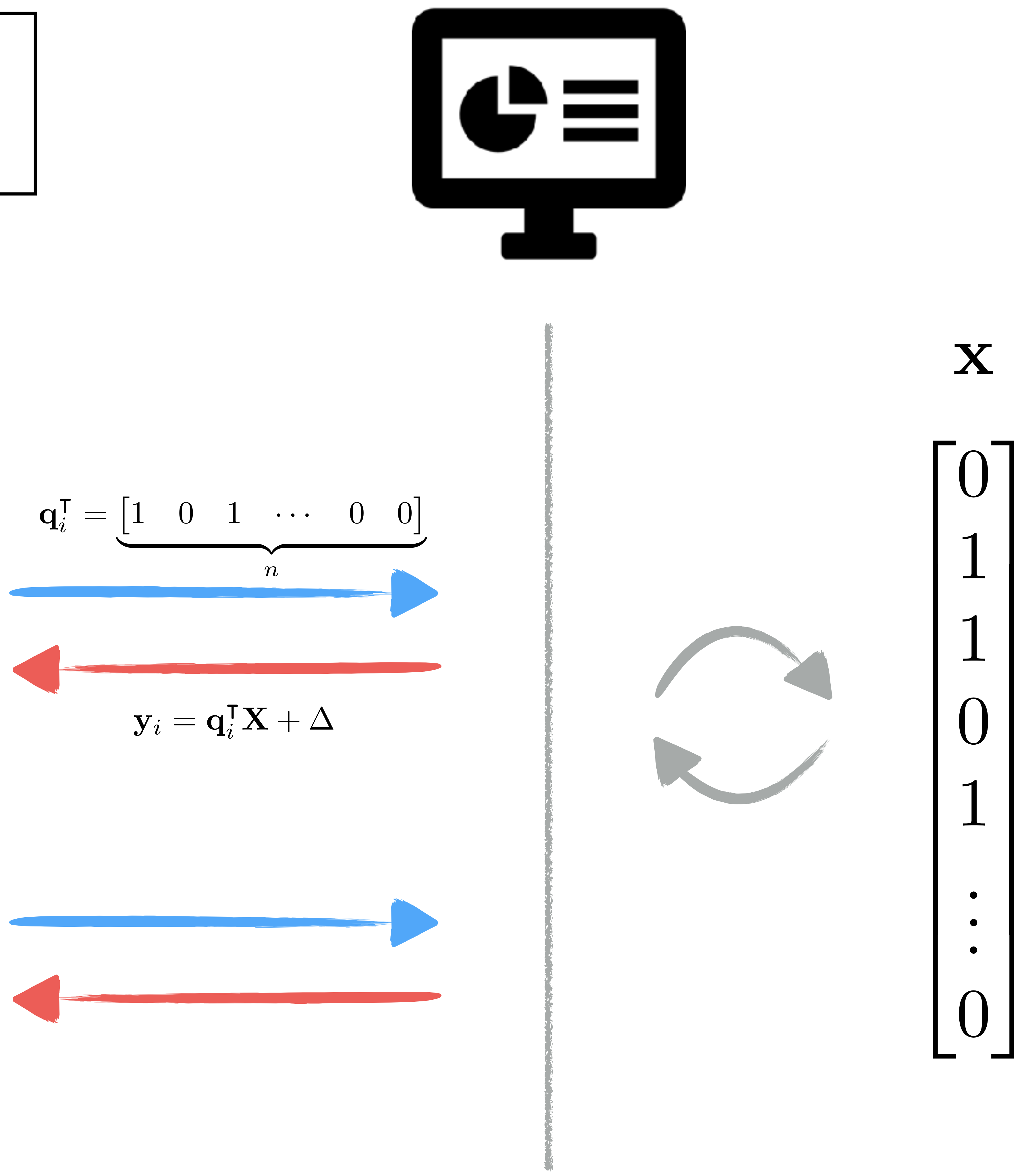




Impossibility of Polynomial Query

at least $\frac{|S_{k_n}|}{\max_i |V_i|}$ queries are required





Impossibility of Polynomial Query

- Therefore, we have the following lower bound on $T_n^*(k_n, \delta_n)$:

$$T_n^*(k_n, \delta_n) \geq \frac{|S_{k_n}|}{\max_{i \in \{1,\ldots,T_n\}} |V_i|}$$

Impossibility of Polynomial Query

- Therefore, we have the following lower bound on $T_n^*(k_n, \delta_n)$:

$$T_n^*(k_n, \delta_n) \geq \frac{|S_{k_n}|}{\max_{i \in \{1,\ldots,T_n\}} |V_i|}$$

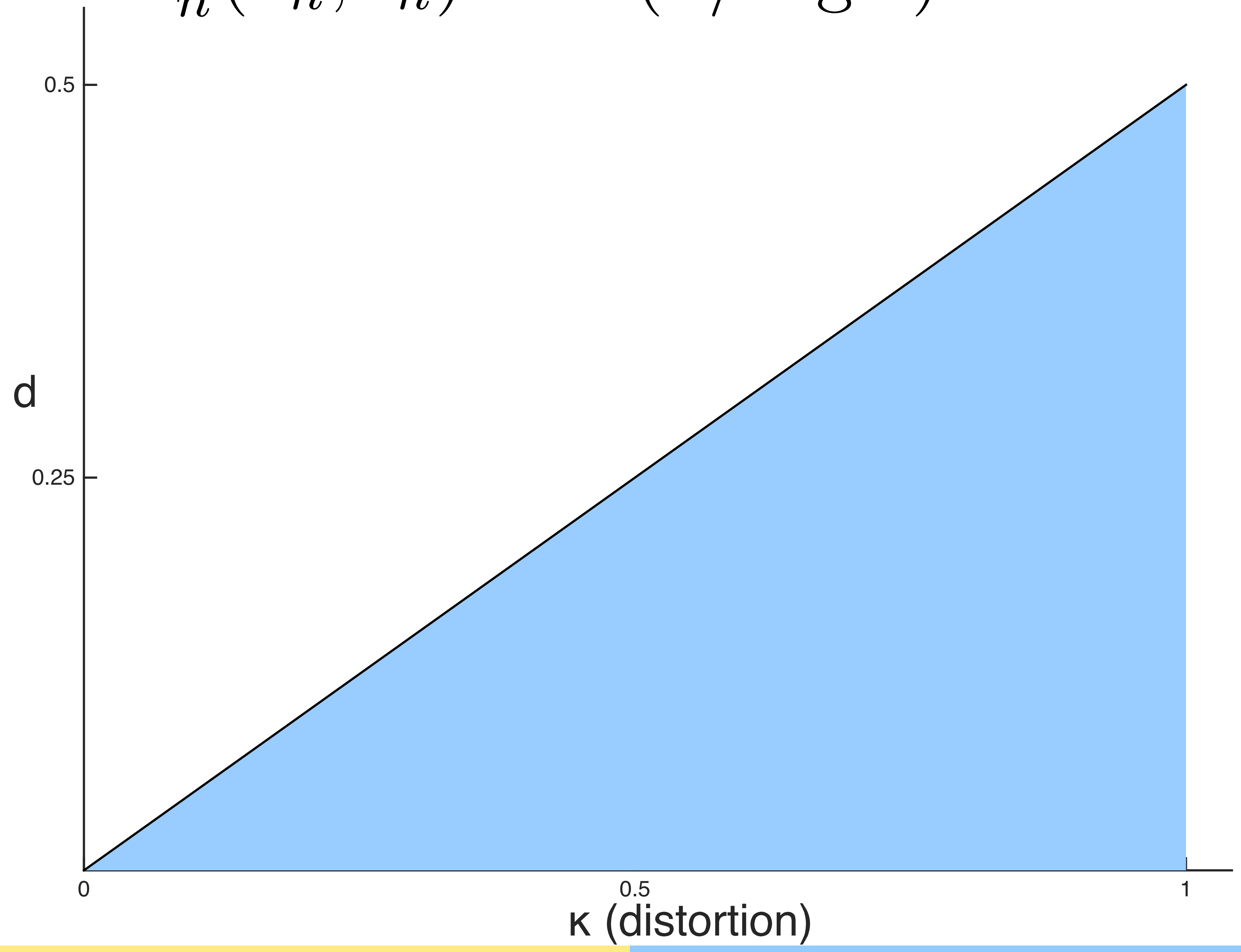
solve the optimization over V, and apply Chernoff ineq.

$$\geq C \exp\left(\frac{\delta_n^2}{k_n}\right) = C \exp\left(n^{2d-\kappa}\right)$$

Regime 2: Achievability and Converse

- Regime 2: $d < \frac{1}{2}\kappa$ (the noise is small enough)

$$T_n^*(k_n, \delta_n) = \Theta\left(n/\log n\right)$$



Achievability : Random Sampling

- Random sampling is considered

Achievability : Random Sampling

- Random sampling is considered
- Evert item is included to the queried subset with probability $1/2$

Achievability : Random Sampling

- Random sampling is considered
- Evert item is included to the queried subset with probability $1/2$

$$\iff (\mathbf{Q})_{i,j} \sim \text{Ber}\left(\frac{1}{2}\right)$$

Achievability : Random Sampling

- Random sampling is considered
- Evert item is included to the queried subset with probability $1/2$

$$\Longleftrightarrow (\mathbf{Q})_{i,j} \sim \mathrm{Ber}\left(\frac{1}{2}\right)$$

- The probability of failure :

$$P_f\left(\mathbf{x}; k_n, \delta_n\right) \triangleq$$
$$P\left\{\exists \text{ a confused } \tilde{\mathbf{x}} \text{ which is consistent with the query output}\right\}$$

Achievability : Random Sampling

- Random sampling is considered
- Evert item is included to the queried subset with probability $1/2$

$$\Longleftrightarrow (\mathbf{Q})_{i,j} \sim \mathrm{Ber}\left(\frac{1}{2}\right)$$

- The probability of failure :

$$P_f\left(\mathbf{x}; k_n, \delta_n\right) \triangleq$$
$$P\left\{\exists \text{ a confused } \tilde{\mathbf{x}} \text{ which is consistent with the query output}\right\}$$

- If number of queries is $\Omega(n/\log n)$ then $P_f(\mathbf{x}; k_n, \delta_n) \to 0$ as $n \to \infty$

Achievability : Random Sampling

- Random sampling is considered
- Evert item is included to the queried subset with probability $1/2$

$$\iff (\mathbf{Q})_{i,j} \sim \mathrm{Ber}\left(\frac{1}{2}\right)$$

- The probability of failure :

$$P_f\left(\mathbf{x}; k_n, \delta_n\right) \triangleq$$
$$P\left\{\exists \text{ a confused } \tilde{\mathbf{x}} \text{ which is consistent with the query output}\right\}$$

- If number of queries is $\Omega(n/\log n)$ then $P_f(\mathbf{x}; k_n, \delta_n) \to 0$ as $n \to \infty$
  - Applying Chernoff bound on failure event

Converse Lower Bound

- Necessary condition :

$$\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X},\ \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 > k_n \implies \|\mathbf{Q}\mathbf{x} - \mathbf{Q}\tilde{\mathbf{x}}\|_\infty > 2\delta_n$$

- Packing inequality :

$$2\delta_n\text{-packing number on } \mathcal{Y} \geq \frac{1}{2}k_n\text{-packing number on } \mathcal{X}$$

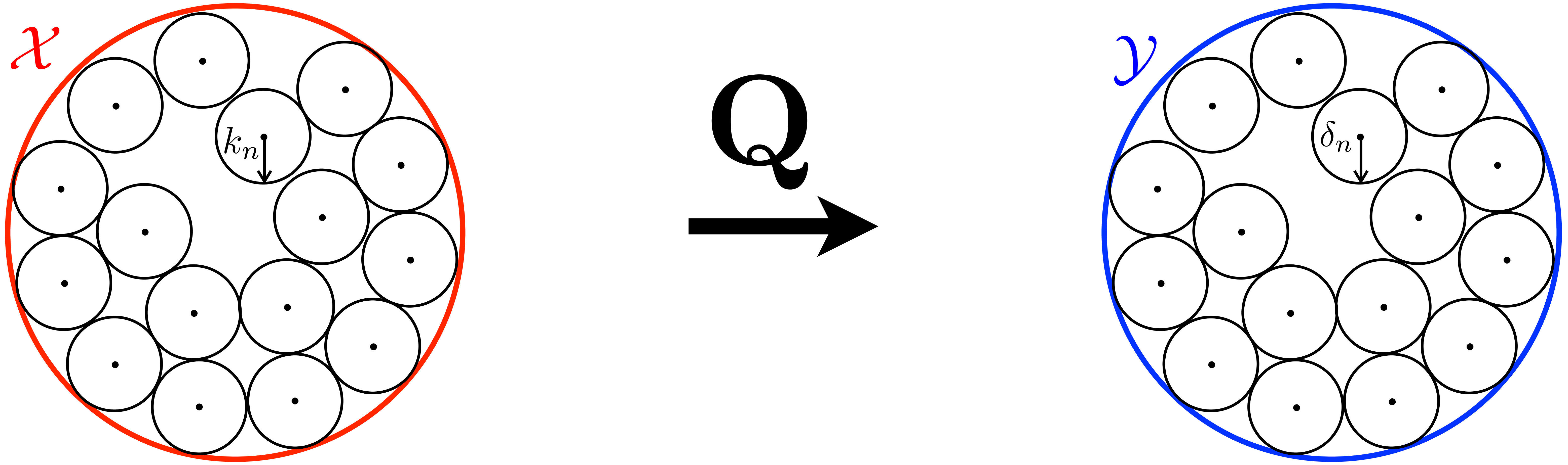
Converse Lower Bound

- Necessary condition :

$$\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X},\ \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 > k_n \implies \|\mathbf{Q}\mathbf{x} - \mathbf{Q}\tilde{\mathbf{x}}\|_\infty > 2\delta_n$$

- Packing inequality :

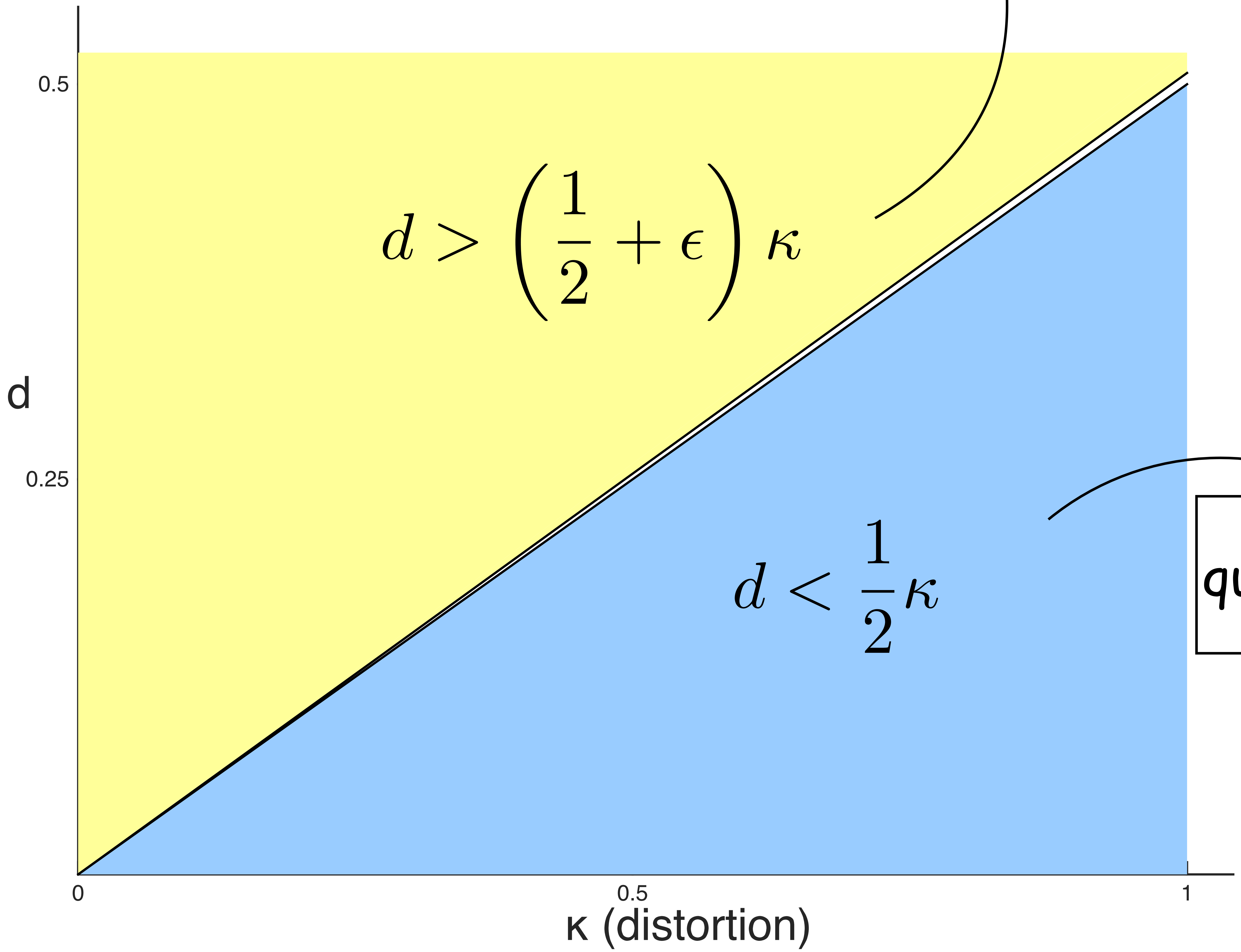
$$2\delta_n\text{-packing number on } \mathcal{Y} \geq \frac{1}{2}k_n\text{-packing number on } \mathcal{X}$$



Summary

$$\delta_n = \Theta\left(n^d\right), \; k_n = \Theta\left(n^\kappa\right)$$

query complexity: non-polynomial $\Omega\left(\exp\left(n^\epsilon\right)\right)$



$$d > \left(\frac{1}{2} + \epsilon\right)\kappa$$

$$d < \frac{1}{2}\kappa$$

query complexity: sub-linear $\Theta\left(\frac{n}{\log n}\right)$

Reference

[1] I.-H. Wang, S.-L. Huang et. al. “Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms,” *Proceedings of IEEE International Symposium on Information Theory*, July 2016.

[2] Ahmed El Alaoui , et. al “Decoding from Pooled Data: Phase Transitions of Message Passing ,” *Proceedings of IEEE International Symposium on Information Theory,* June 2017

[3] C. Dwork, A. Roth, “The algorithmic foundations of differential privacy,” *Theoretical Computer Science,* 2013

Question ?

Thank you for your attention !