

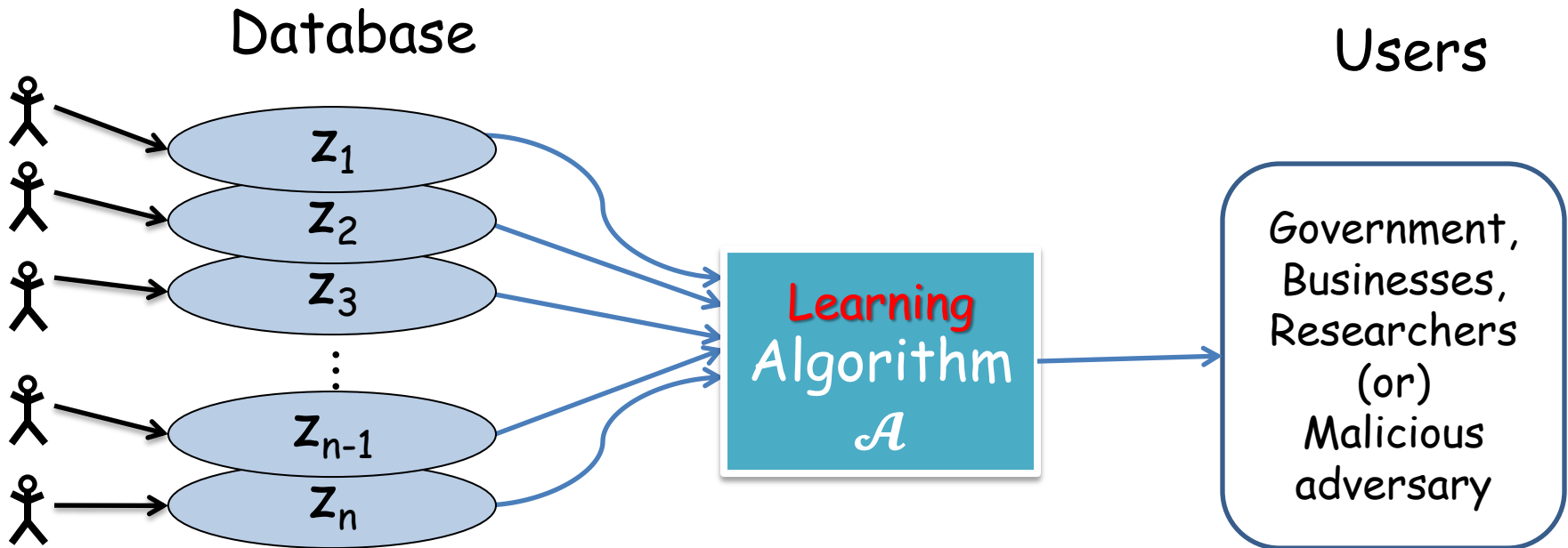
Private Learning and Sanitization: Pure vs. Approx. Differential Privacy

Uri Stemmer

Ben-Gurion University

Join work with Amos Beimel and Kobbi Nissim

Why Private Learners?



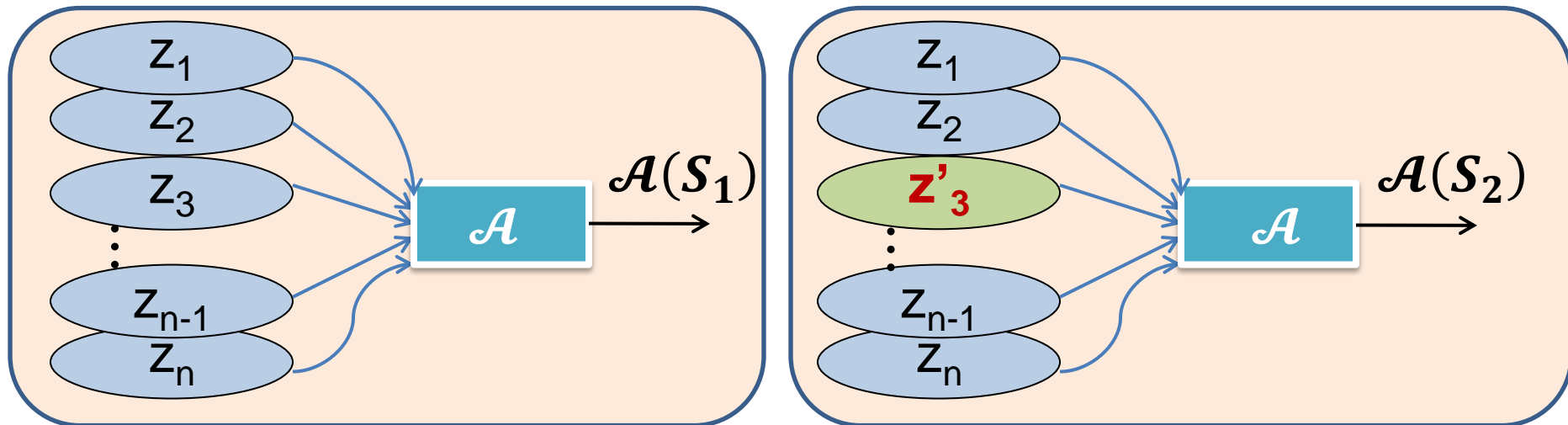
Often, this algorithmic task can be abstracted as a learning problem:

- Bank is interested in predicting (based on past customers) whether new customers are good/bad credit

Differential Privacy

Dwork, McSherry, Nissim, Smith 2006

Changing one record does not change the output distribution "too much"



Differential Privacy

Dwork, McSherry, Nissim, Smith 2006

Changing one record does not change the output distribution “too much”

A (rand) algorithm \mathcal{A} is differentially private if for all neighboring databases S_1, S_2 and for all sets of outputs F :

$$\Pr[\mathcal{A}(S_1) \in F] \approx \Pr[\mathcal{A}(S_2) \in F]$$

Pure Differential Privacy

Dwork, McSherry, Nissim, Smith 2006

Changing one record does not change the output distribution “too much”

A (rand) algorithm \mathcal{A} is ϵ differentially private if for all neighboring databases S_1, S_2 and for all sets of outputs F :

$$\Pr[\mathcal{A}(S_1) \in F] \leq e^\epsilon \cdot \Pr[\mathcal{A}(S_2) \in F]$$

Approx. Differential Privacy

Dwork, McSherry, Nissim, Smith 2006

Dwork, Kenthapadi, McSherry, Mironov, Naor 2006

Changing one record does not change the output distribution "too much"

A (rand) algorithm \mathcal{A} is (ϵ, δ) differentially private if for all neighboring databases S_1, S_2 and for all sets of outputs F :

$$\Pr[\mathcal{A}(S_1) \in F] \leq e^\epsilon \cdot \Pr[\mathcal{A}(S_2) \in F] + \delta$$

Our Results:

- Sample complexity of **Private Learning** and **Sanitization** can be drastically smaller if we settle for **approximate** differential privacy.
- **Label Privacy** [Chaudhuri and Hsu 2011]
Learning model with weakened privacy demands.
We settle the question of sample complexity: **$O(VC)$** .
 - Same as non-private learning.
 - Not in this talk.
- Natural connection between Private Learning and Sanitization, leads to lower bounds on Sanitization.
 - Not in this talk.

What is Private Learning?

Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith 08

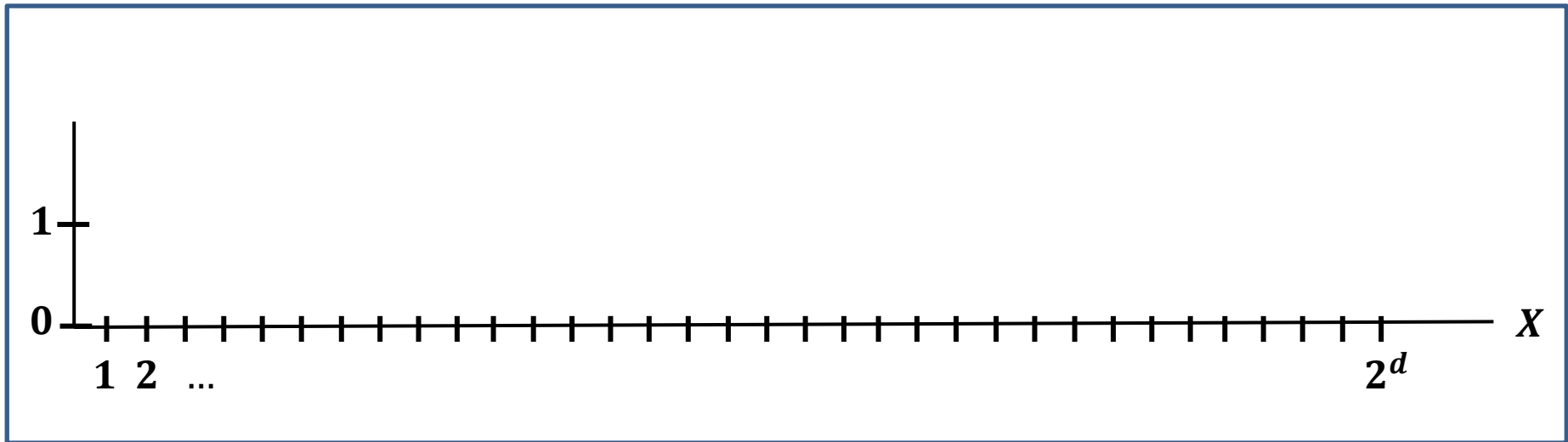
Definition:

+ PAC Learning
+ Differential Privacy

Private Learning

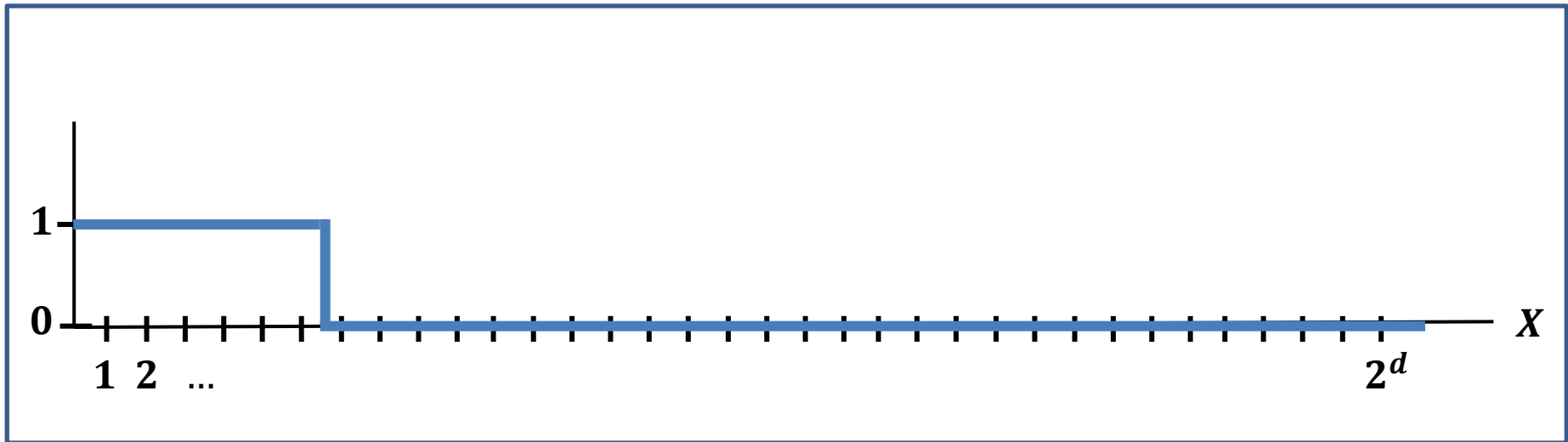
“PAC” Model [Valiant 84]

- Domain X .



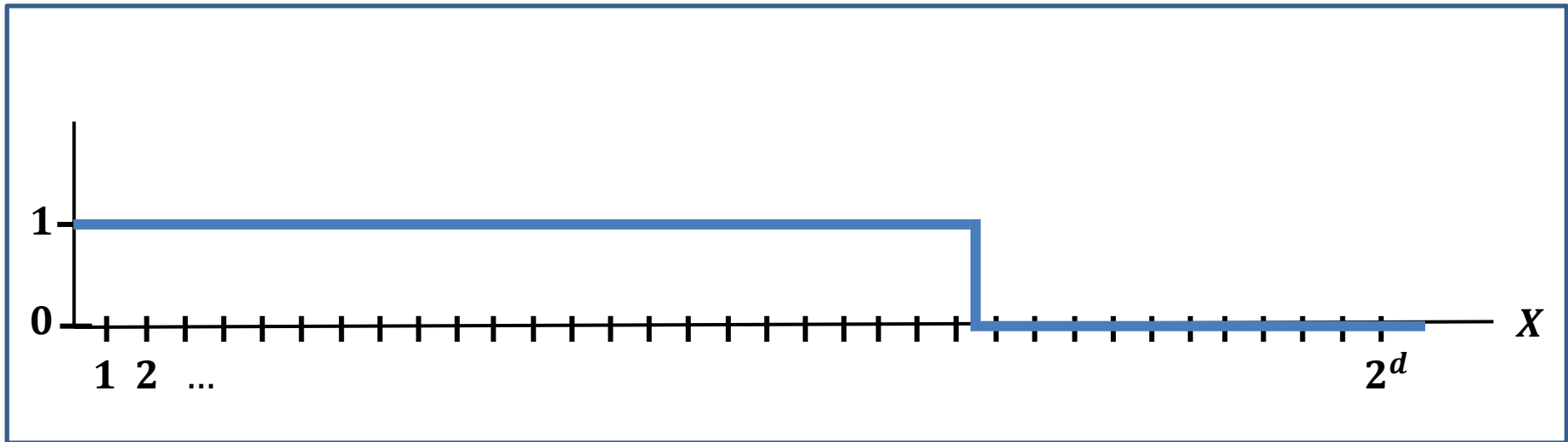
“PAC” Model [Valiant 84]

- Domain X .
- Set \mathcal{C} of boolean functions over X .
 - for example: INTERVAL_d



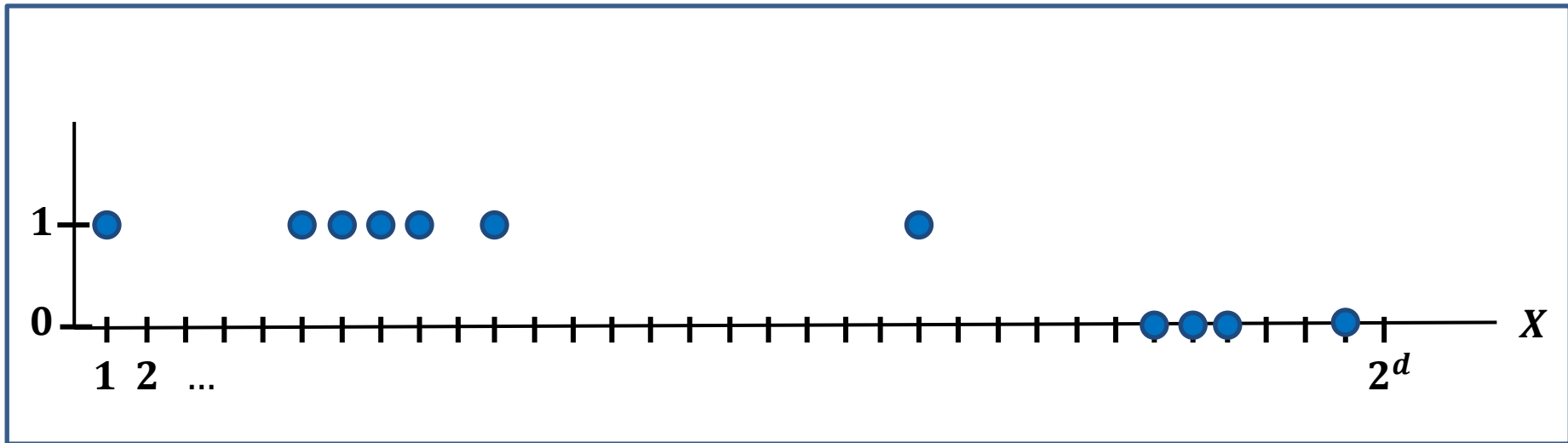
“PAC” Model [Valiant 84]

- Domain X .
- Set \mathcal{C} of boolean functions over X .
 - for example: INTERVAL_d



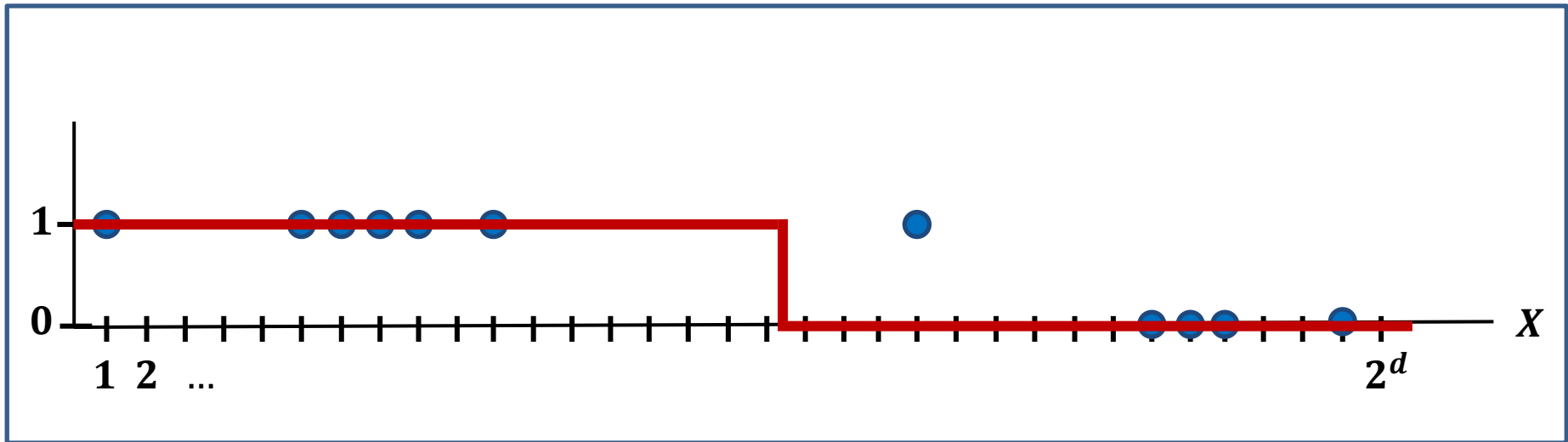
“PAC” Model [Valiant 84]

- Domain X .
- Set \mathcal{C} of boolean functions over X .
 - for example: INTERVAL_d
- Labeled sample.



“PAC” Model [Valiant 84]

- Domain X .
- Set \mathcal{C} of boolean functions over X .
 - for example: INTERVAL_d
- Labeled sample.
- **Output a classifier h .**



Related work in Private Learning (partial list)

[BDMN 05] First private learning algorithms. SQ based.

[KLNRS 08] Define private learning, and showed:
Every class \mathcal{C} can be privately learned using $\log|\mathcal{C}|$ labeled samples.

[BKN 10] Sample complexity of private learning.

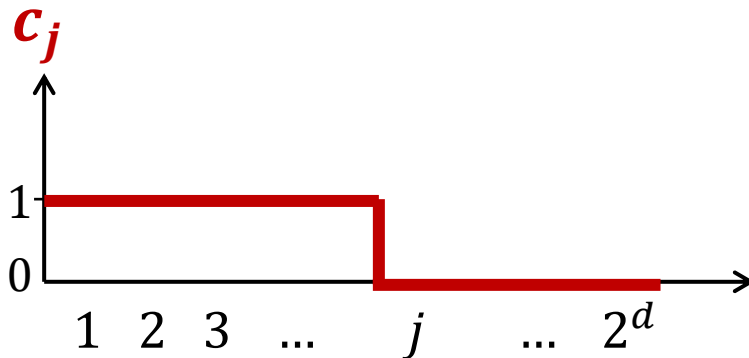
[CH 11] Learning in continuous domain, label privacy.

[CM 08, CMS 11, KST 12] Machine learning.

[BLR 08, DNRRV 09, ...] Synthetic Data.

[DRV 10] Private Boosting.

Running Example: INTERVAL_d

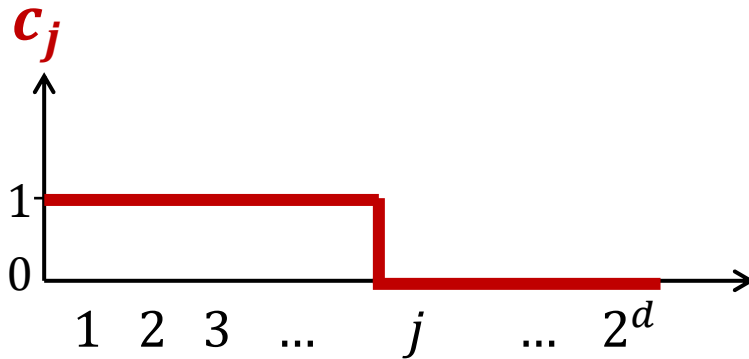


$$c_j(x) = 1 \Leftrightarrow x < j$$

Facts:

- non-private **proper** learner with $O(1)$ samples.
- ϵ -private **proper** learner: $\Theta(d)$ samples [BBKN 10].

Running Example: INTERVAL_d



$$c_j(x) = 1 \Leftrightarrow x < j$$

Facts:

- non-private **proper** learner with $O(1)$ samples.
- ϵ -private **proper** learner: $\Theta(d)$ samples [BBKN 10].

We show:

(ϵ, δ) -private **proper** learner with $2^{O(\log^* d)}$ samples.

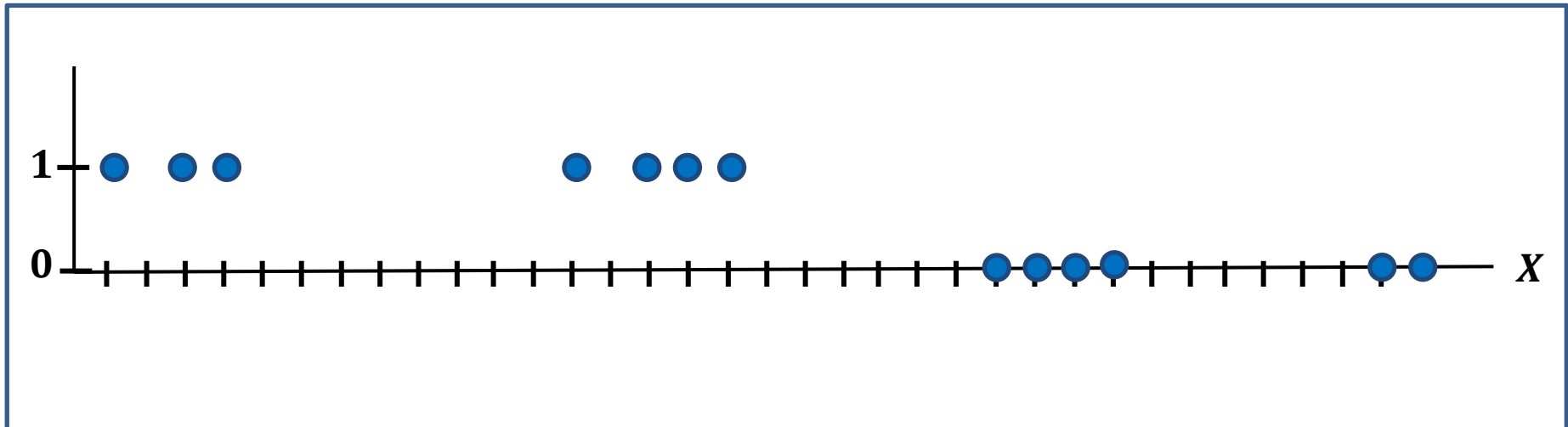
Privately Learning intervals: Ideas and Intuition.

We show:

(ϵ, δ) -private **proper** learner with $2^{O(\log^* d)}$ samples.

The Goal:

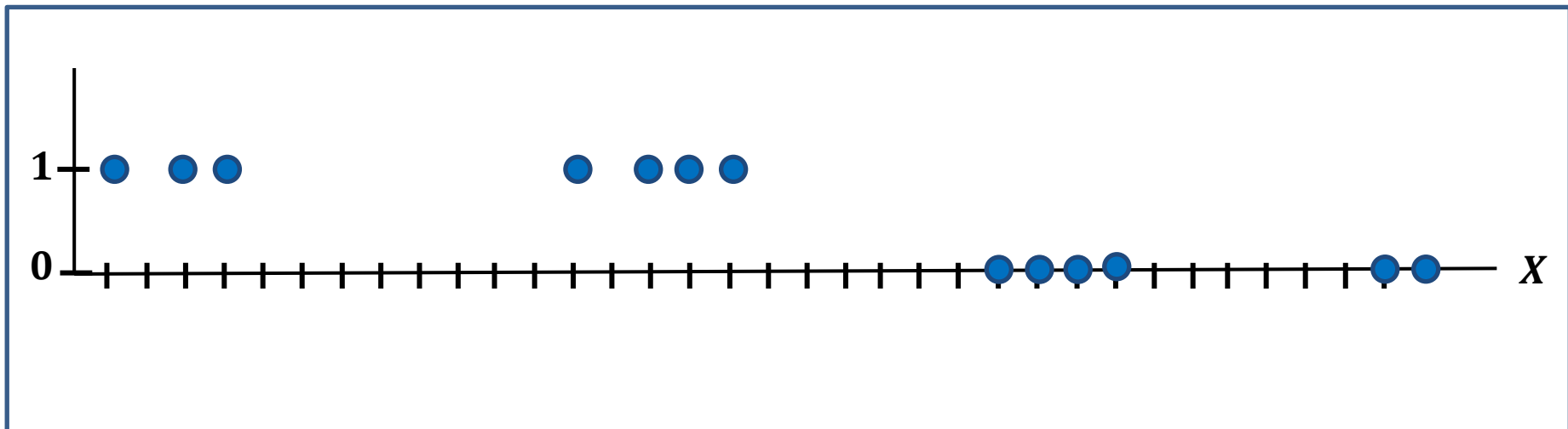
Given a labeled sample, choose a concept with small error.



4-good interval G

Assume we can (privately) obtain an interval $G \subseteq X$ s.t.

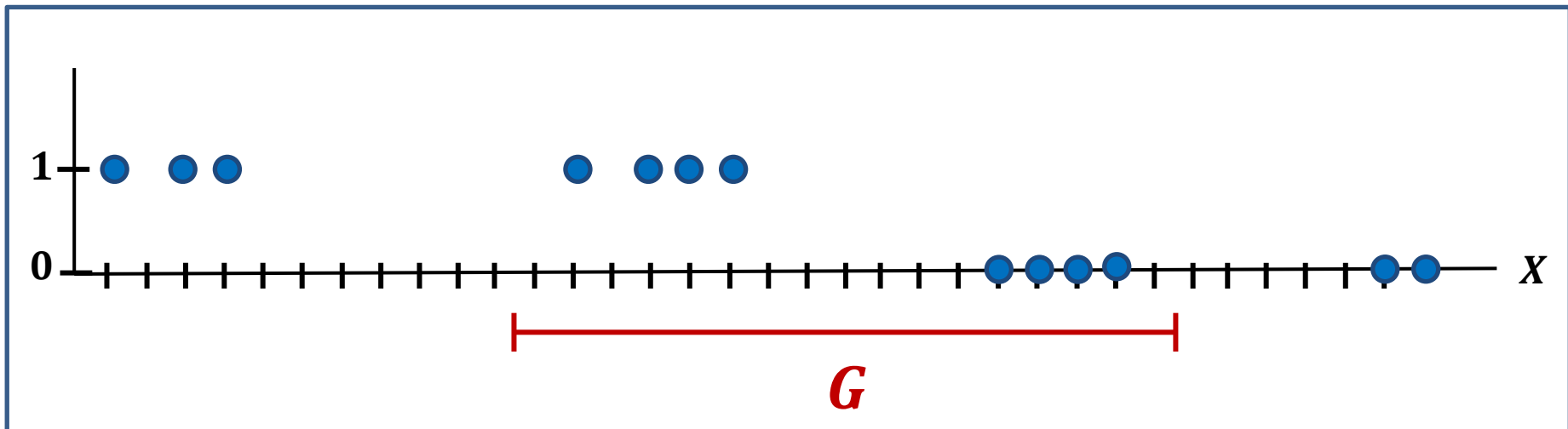
- **Contains** "a lot" of ones, and "a lot" of zeroes.
- Every interval $I \subseteq X$ of **length** $\leq |G|/4$ either does **not contain** "too many" ones or does **not contain** "too many" zeroes.



4-good interval G

Assume we can (privately) obtain an interval $G \subseteq X$ s.t.

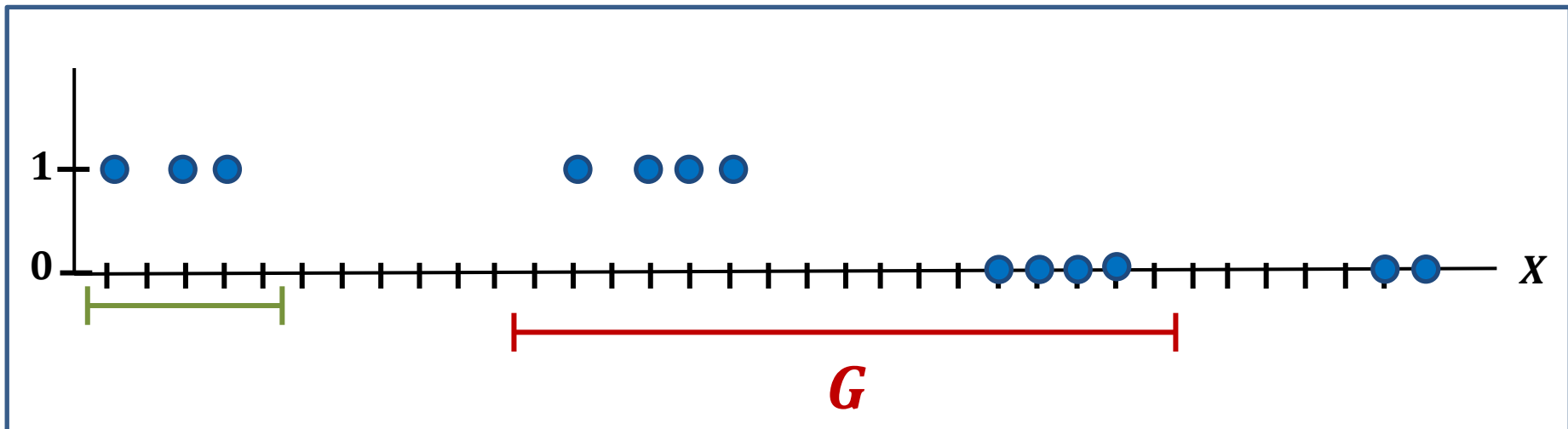
- **Contains** "a lot" of ones, and "a lot" of zeroes.
- Every interval $I \subseteq X$ of **length** $\leq |G|/4$ either does **not contain** "too many" ones or does **not contain** "too many" zeroes.



4-good interval G

Assume we can (privately) obtain an interval $G \subseteq X$ s.t.

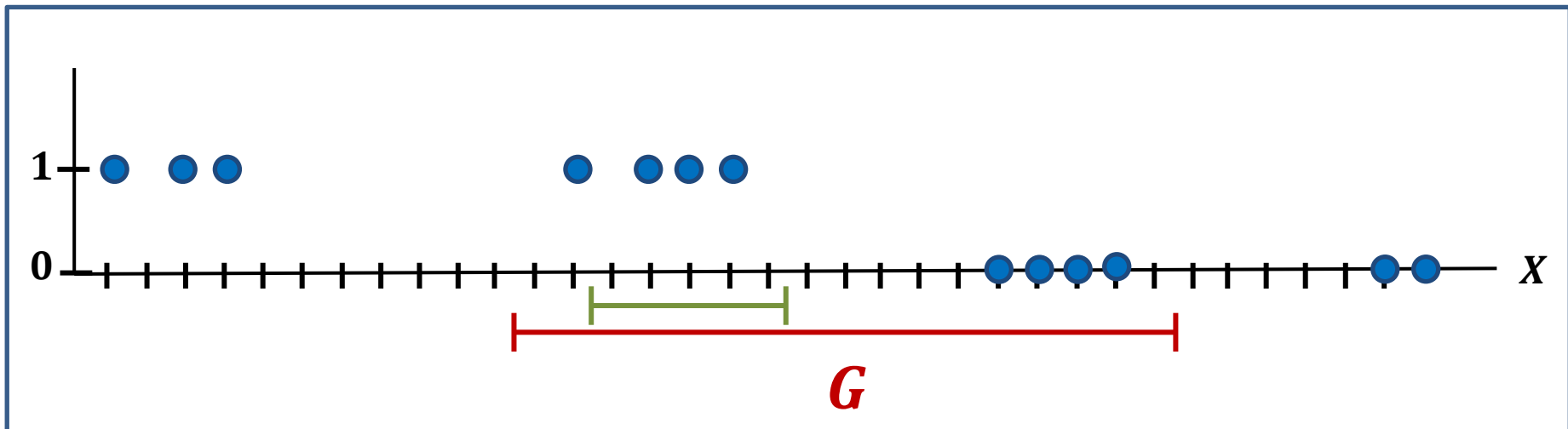
- **Contains** "a lot" of ones, and "a lot" of zeroes.
- Every interval $I \subseteq X$ of **length** $\leq |G|/4$ either does **not contain** "too many" ones or does **not contain** "too many" zeroes.



4-good interval G

Assume we can (privately) obtain an interval $G \subseteq X$ s.t.

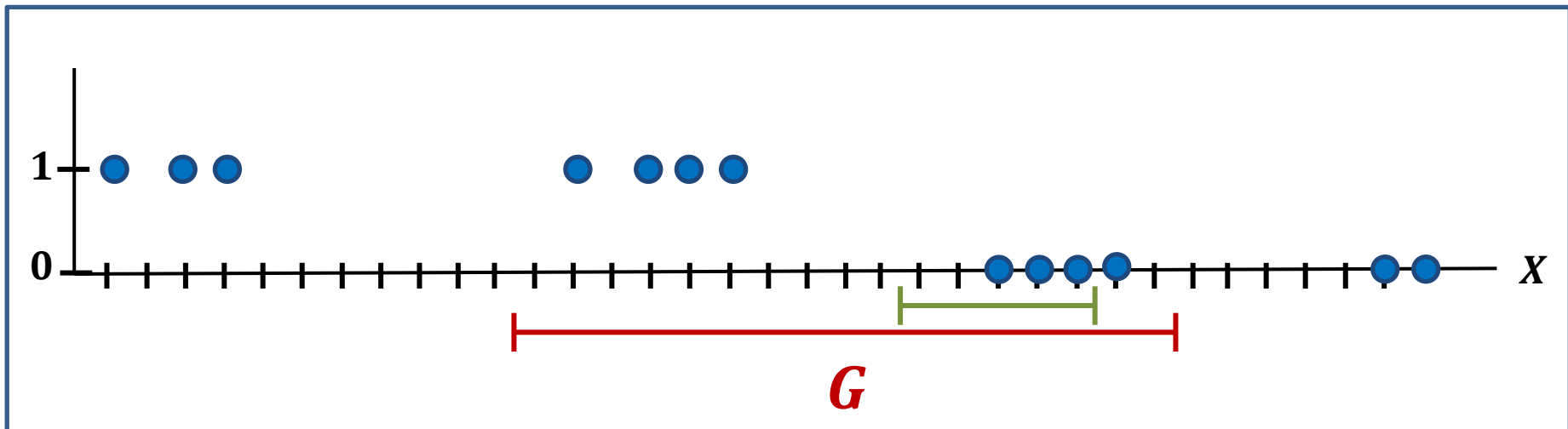
- **Contains** "a lot" of ones, and "a lot" of zeroes.
- Every interval $I \subseteq X$ of **length** $\leq |G|/4$ either does **not contain** "too many" ones or does **not contain** "too many" zeroes.



4-good interval G

Assume we can (privately) obtain an interval $G \subseteq X$ s.t.

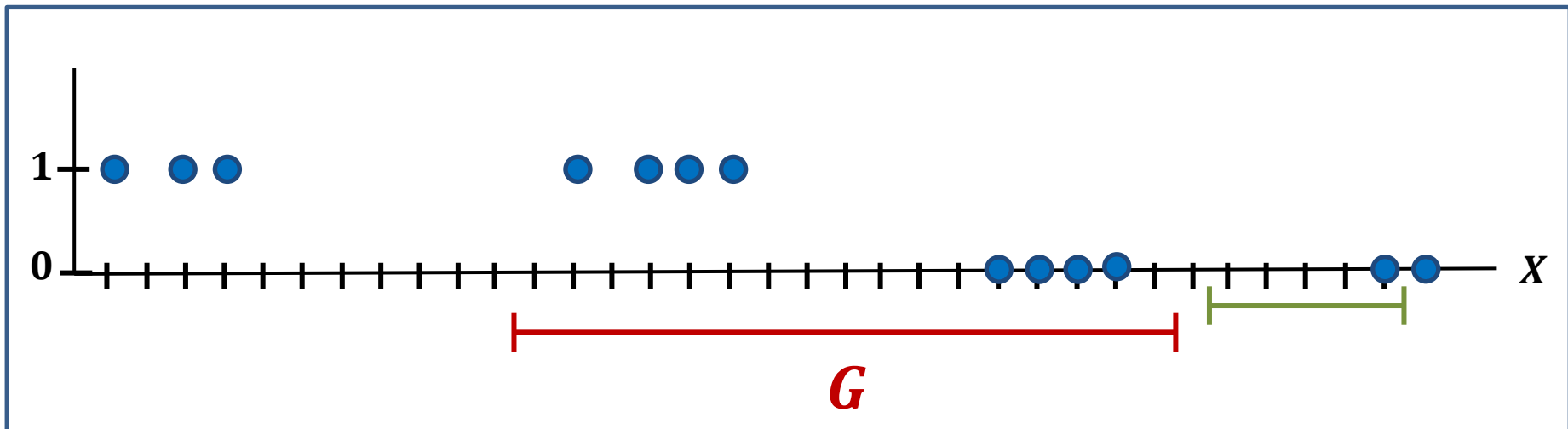
- **Contains** "a lot" of ones, and "a lot" of zeroes.
- Every interval $I \subseteq X$ of **length** $\leq |G|/4$ either does **not contain** "too many" ones or does **not contain** "too many" zeroes.



4-good interval G

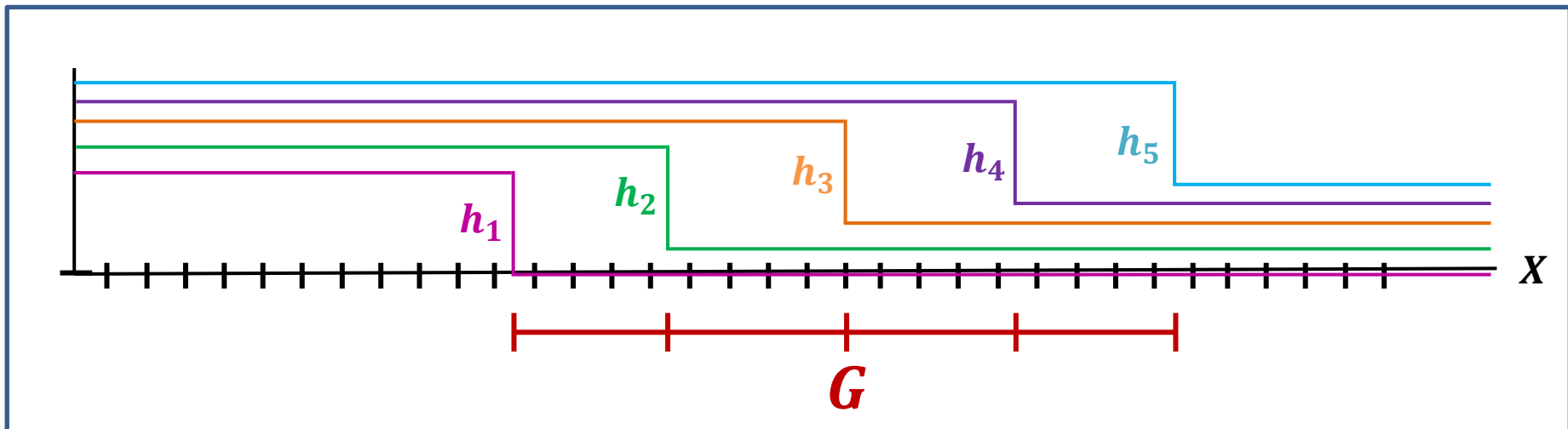
Assume we can (privately) obtain an interval $G \subseteq X$ s.t.

- **Contains** "a lot" of ones, and "a lot" of zeroes.
- Every interval $I \subseteq X$ of **length** $\leq |G|/4$ either does **not contain** "too many" ones or does **not contain** "too many" zeroes.



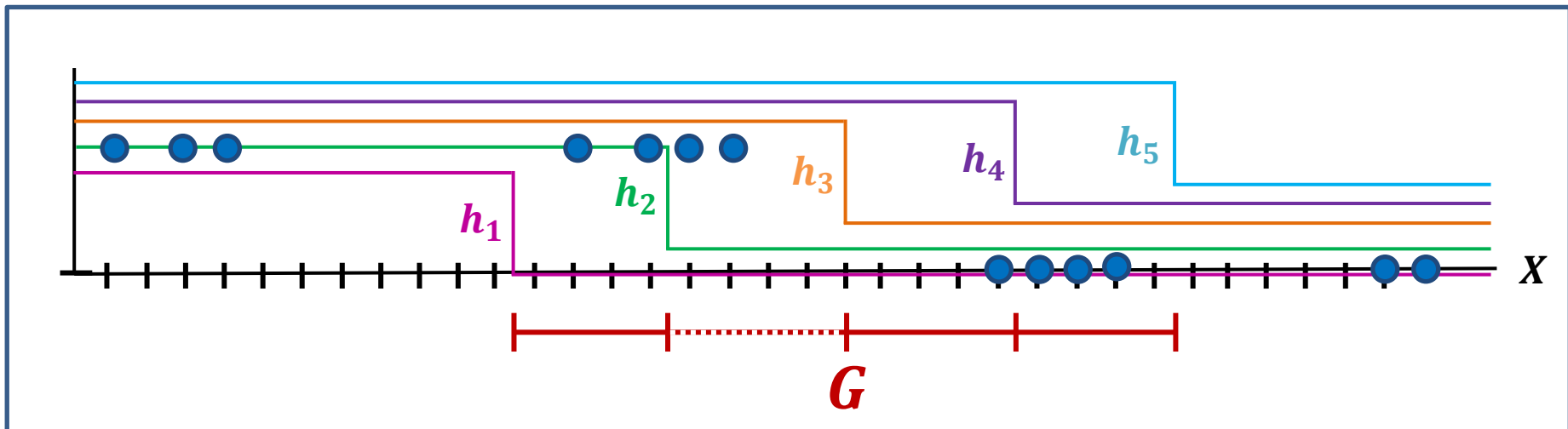
4-good interval $G \Rightarrow$ done!

- Divide G into 4 equal intervals, and define 5 “equally spread” concepts in G .
- At least one concept has small error.



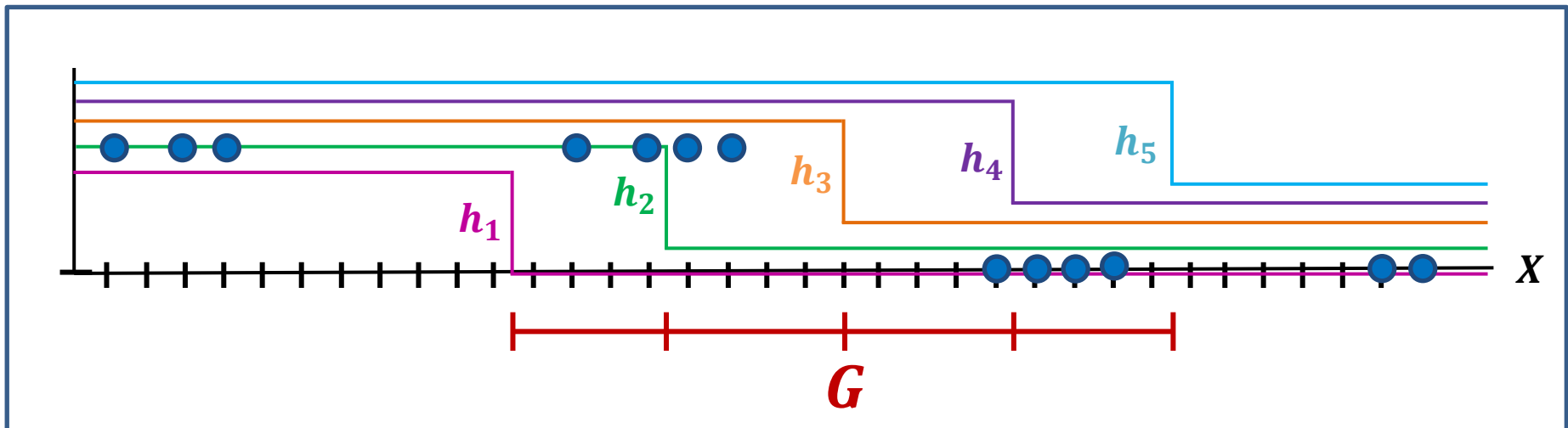
4-good interval $G \Rightarrow$ done!

- Divide G into 4 equal intervals, and define 5 “equally spread” concepts in G .
- At least one concept has small error.



4-good interval $G \Rightarrow$ done!

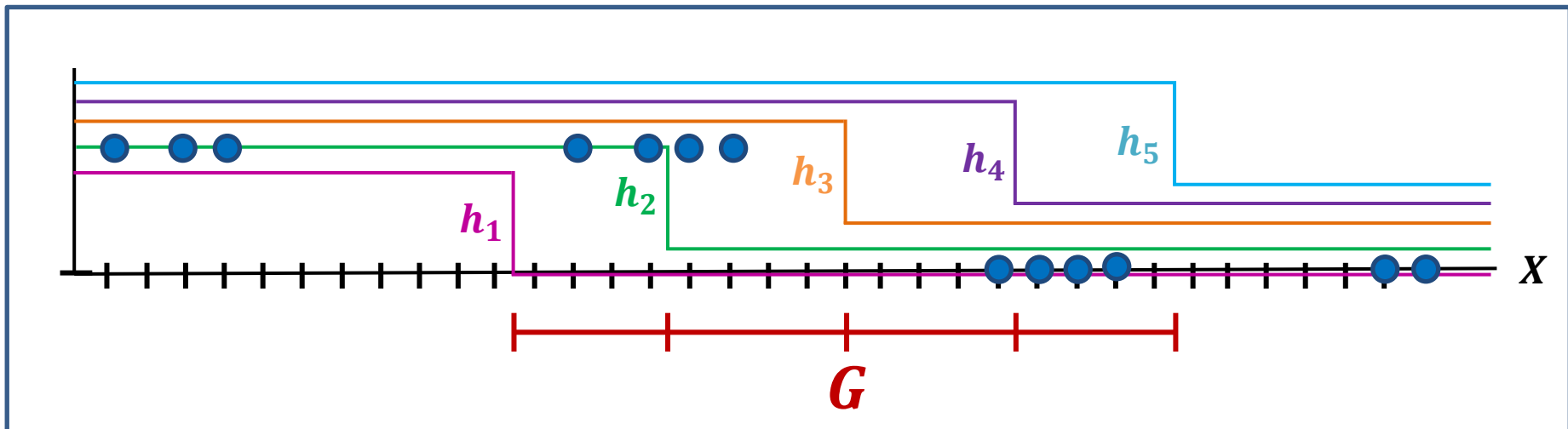
- Divide G into 4 equal intervals, and define 5 “equally spread” concepts in G .
- At least one concept has small error.
- Choose one using the Exp. Mechanism [McSherry and Talwar 07] (requires $O(1)$ samples).



4-good interval $G \Rightarrow$ done!

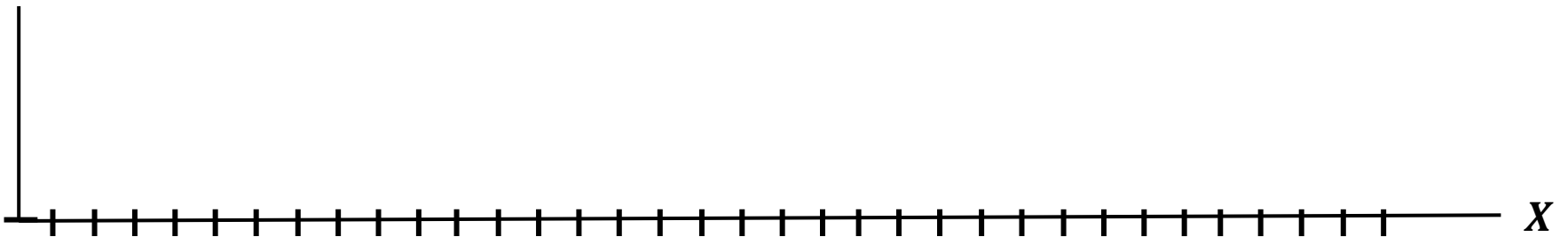
- Divide G into 4 equal intervals, and define 5 “equally spread” concepts in G .
- At least one concept has small error.
- Choose one using the Exp. Mechanism [McSherry and Talwar 07] (requires $O(1)$ samples).

Conclusion: suffices to find a 4-good interval.



Finding a 4-good interval

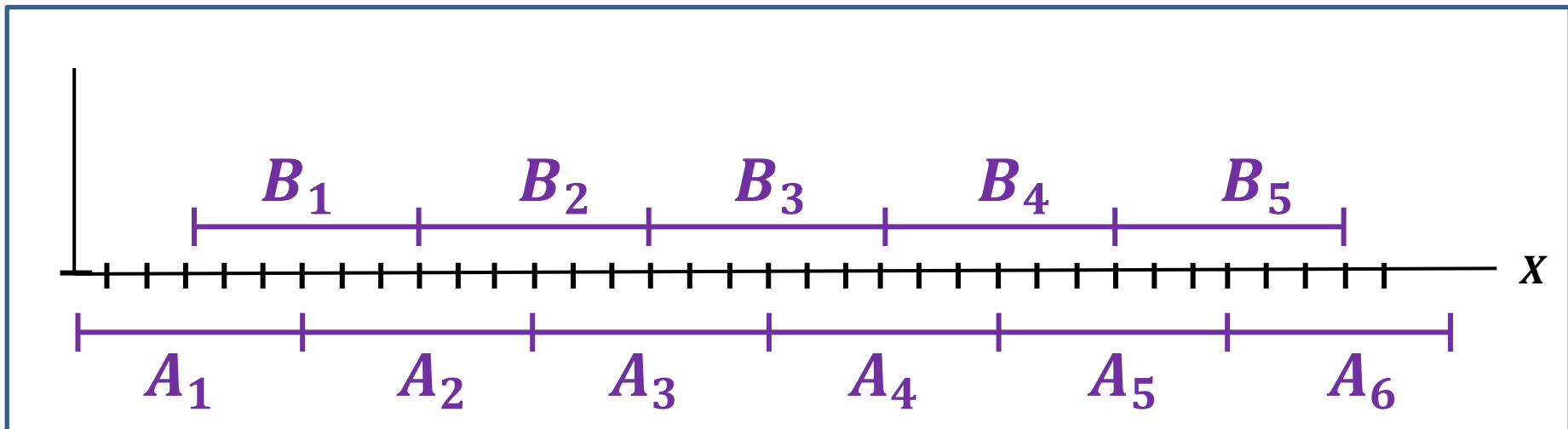
Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

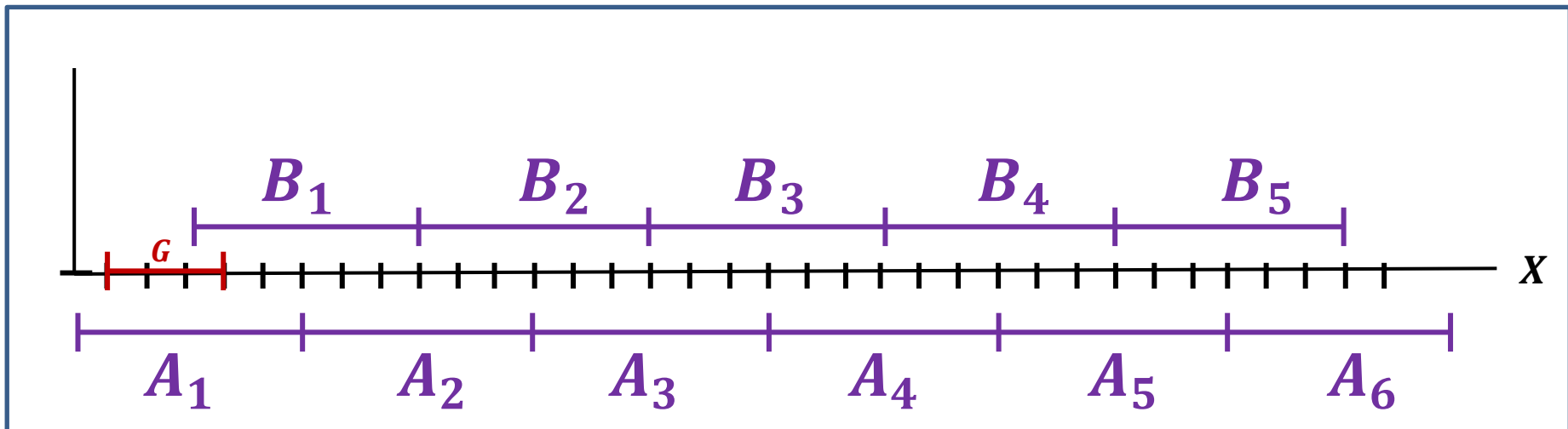
- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

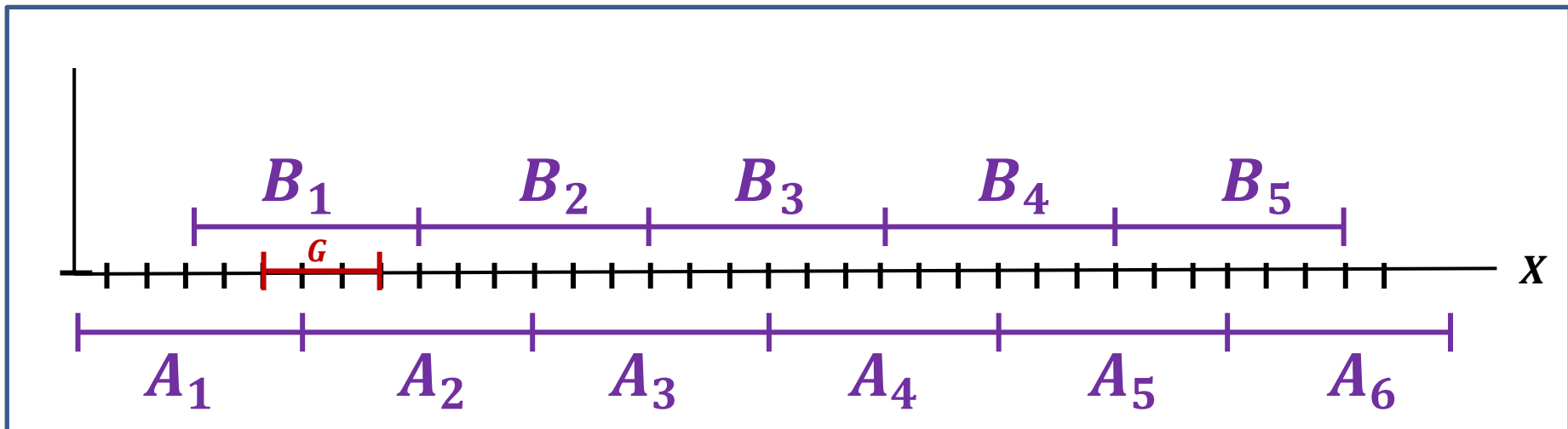
- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

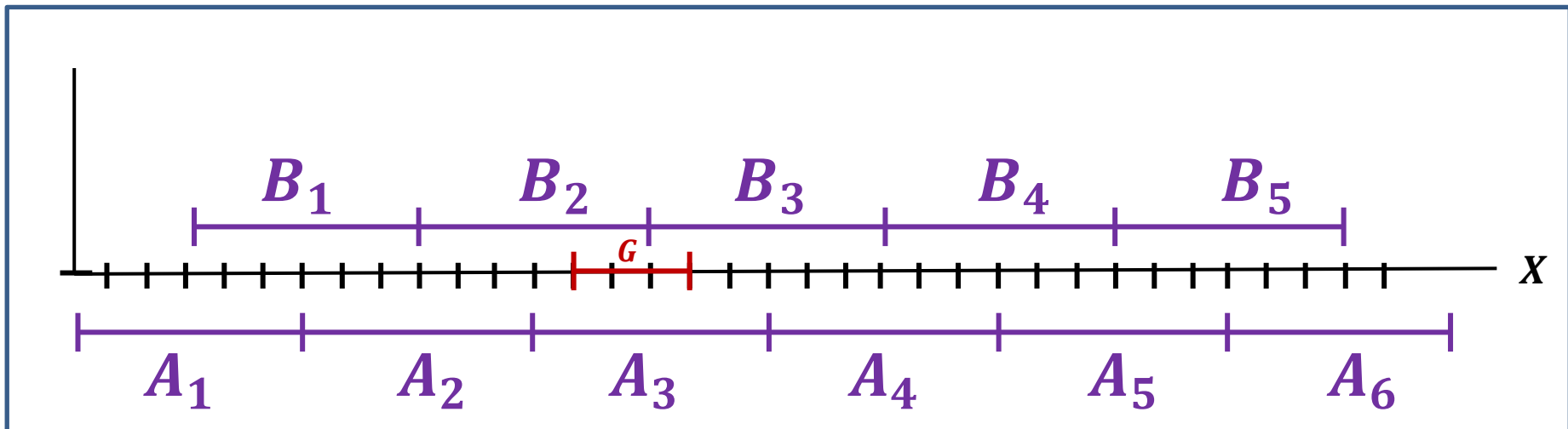
- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

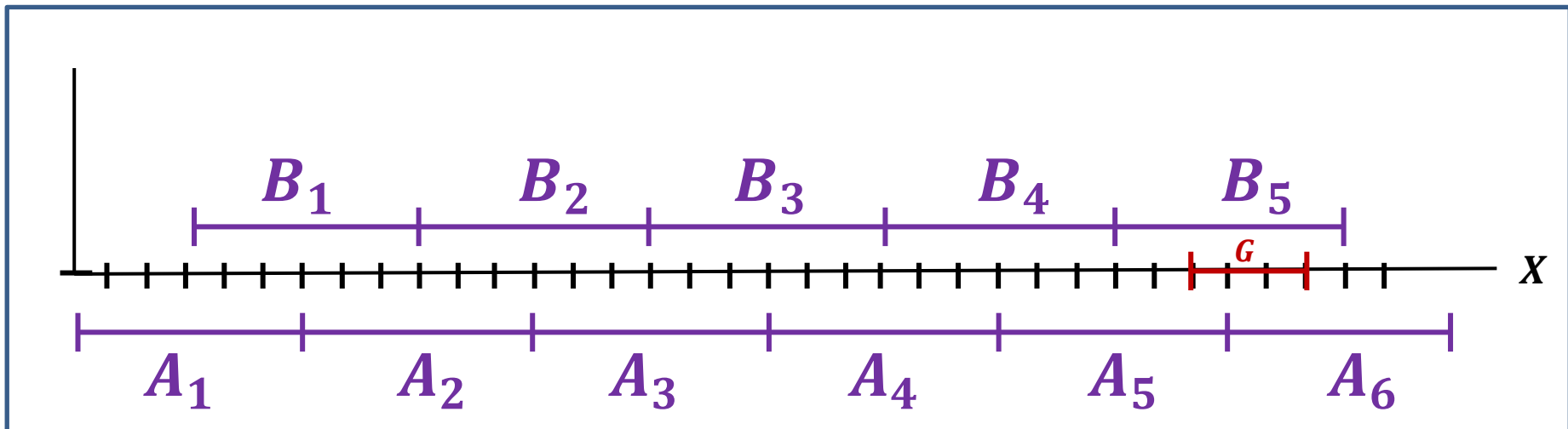
- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .

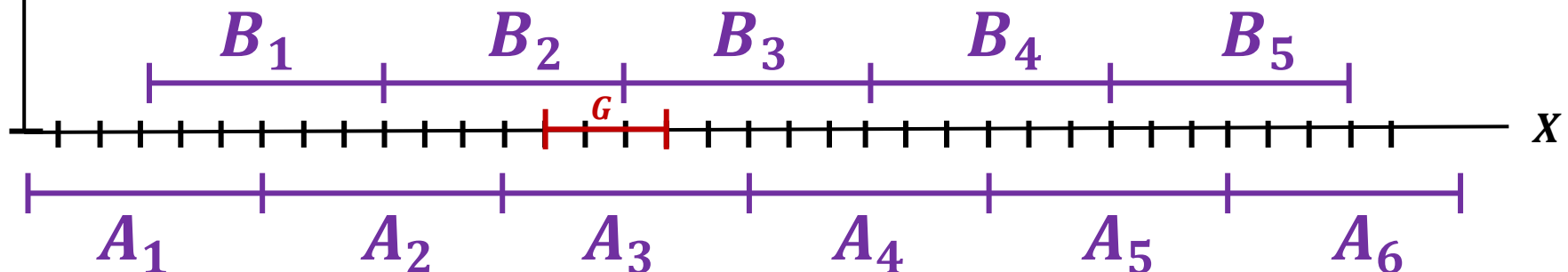


Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .

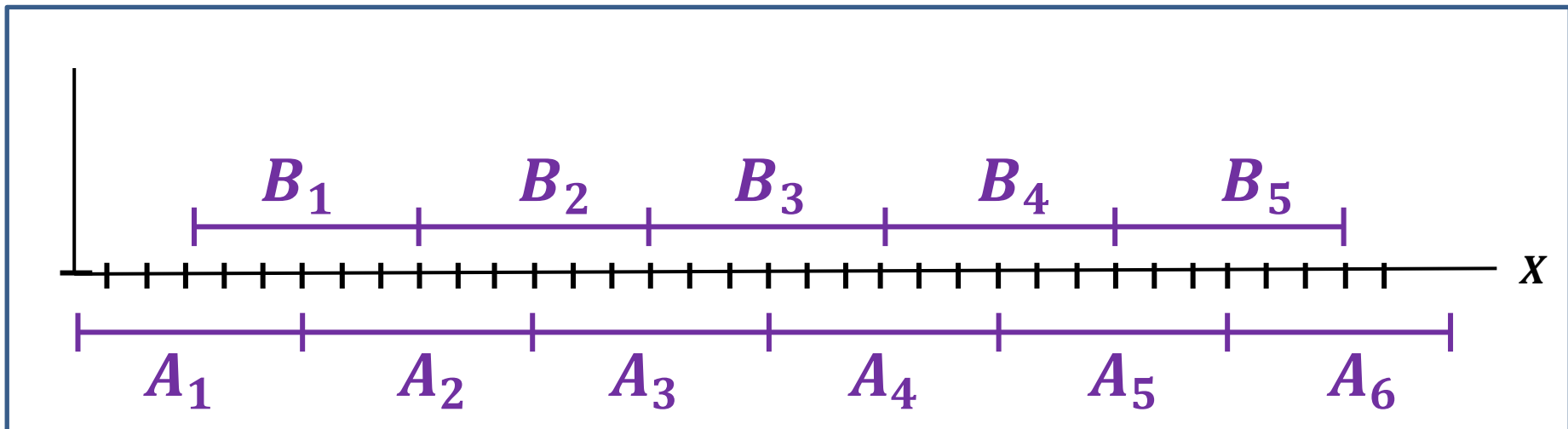
- Say $G \in A_3$. Then A_3 contains "lots" of ones and zeroes.
- Every other A_i cannot contain both ones and zeroes.
- Look for A_i with "lots" of ones and zeroes.



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

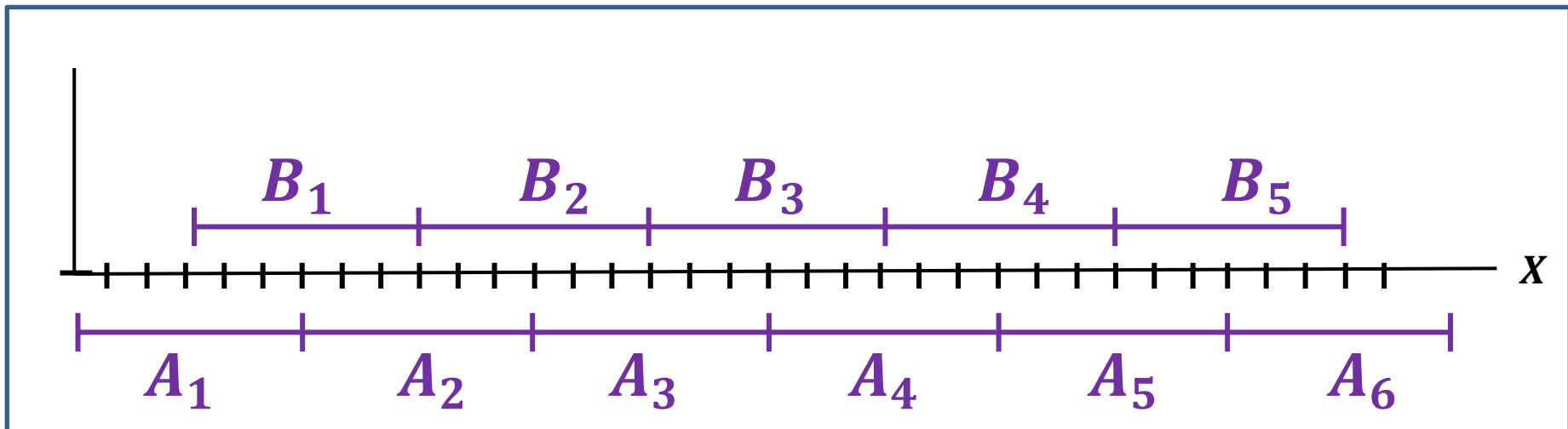
- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

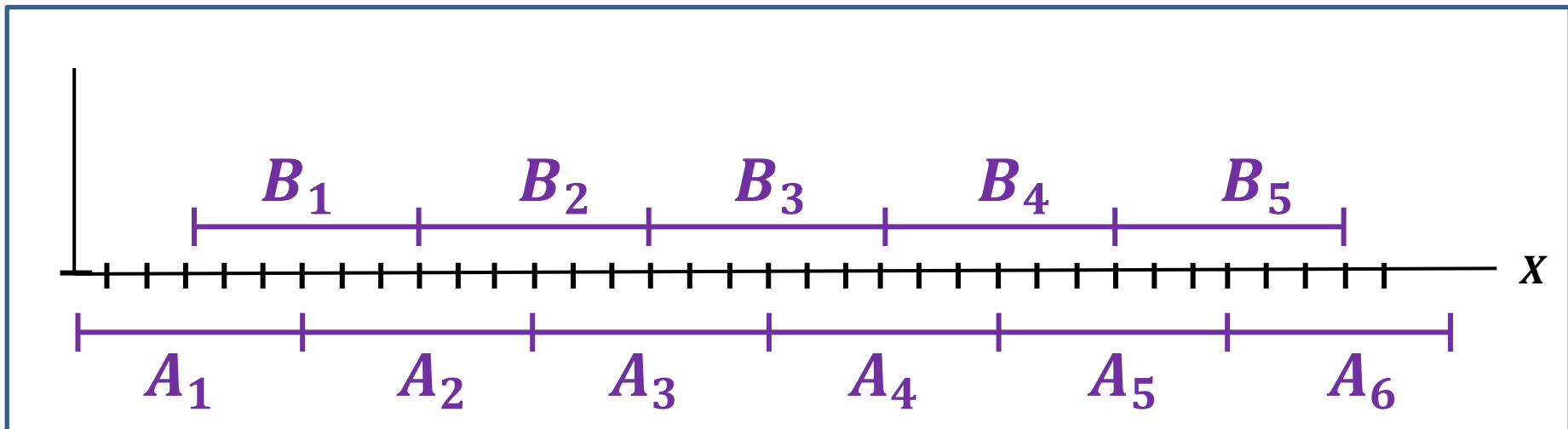
- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .
- Choose an interval using A_{dist} [ST 2013] (requires $O(1)$ samples).



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .
- Choose an interval using A_{dist} [ST 2013] (requires $O(1)$ samples).
- The chosen interval is of length $2|G| \implies 4\text{-good!}$



Finding a 4-good interval

Assume we can (privately) obtain a $J \in \mathbb{R}$ s.t. there exists a 2-good interval G of length J .

- Divide X into intervals $\{A_i\}$ and $\{B_i\}$ of length $2J$, where the $\{B_i\}$'s are right-shifted by J .
- At least one interval contains G .
- Choose an interval using A_{dist} [ST 2013] (requires $O(1)$ samples).
- The chosen interval is of length $2|G| \implies 4\text{-good!}$

Conclusion: suffices to find a length J of a 2-good interval.

A_1

A_2

A_3

A_4

A_5

A_6

Computing the length J

Easy solution:

- Noisy binary search on $0 \leq J \leq 2^d$.
- d noisy comparisons requires d samples.

Computing the length J

Easy solution:

- Noisy binary search on $0 \leq J \leq 2^d$.
- d noisy comparisons requires d samples.

Better solution:

- Noisy binary search on the power $0 \leq j \leq \log d$ of a 2-good interval of length $J = 2^j$.
- $\log d$ noisy comparisons requires $\log d$ samples.

Computing the length J

Easy solution:

- Noisy binary search on $0 \leq J \leq 2^d$.
- d noisy comparisons requires d samples.

Better solution:

- Noisy binary search on the power $0 \leq j \leq \log d$ of a 2-good interval of length $J = 2^j$.
- $\log d$ noisy comparisons requires $\log d$ samples.

In the paper:

Use recursion on binary search and significantly reduce the costs.

Theorem:

There exists an (ϵ, δ) -private learner for INTERVAL_d with sample complexity $2^{\log^* d}$.

Summary and Open Problems

- **What we saw:**

Efficient (ϵ, δ) -private learner for INTERVAL_d with low sample complexity.

- This **separates** the sample complexity of (ϵ, δ) -private and ϵ -private learners.

- **Other results:**

- Efficient (ϵ, δ) -private for other concept classes with **even lower** sample complexity (**independent of the domain**).

- Similar results for Data Sanitization.

- **Open problem:**

Lower bounds on the sample complexity of (ϵ, δ) -private learners?