Rectangle Inequalities for Data Structure Lower Bounds

Yitong Yin
Nanjing University

Nexus of Information and Computation Theories
Fundamental Inequalities and Lower Bounds Theme
@ Institut Henri Poincaré

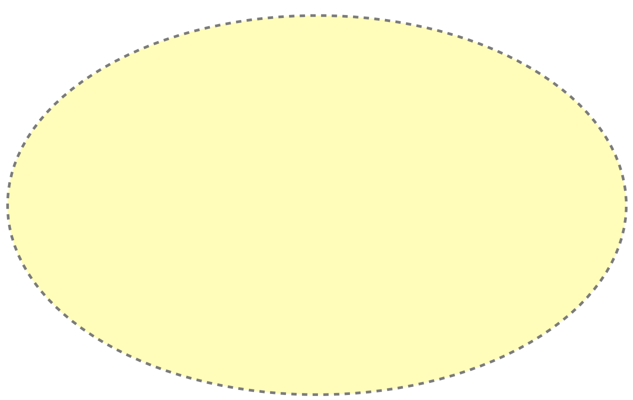
Online Note

"Yitong Yin: Simple average-case lower bounds for approximate near-neighbor from isoperimetric inequalities."

http://arxiv.org/abs/1602.05391

Nearest Neighbor Search
(*NNS*)

metric space (X, dist)

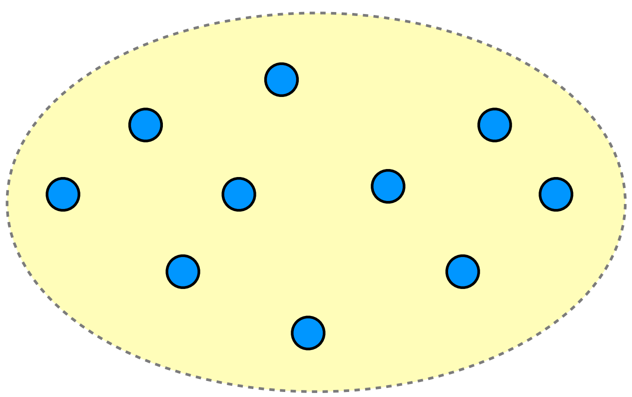


Nearest Neighbor Search
(*NNS*)

metric space (X,dist)

database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$



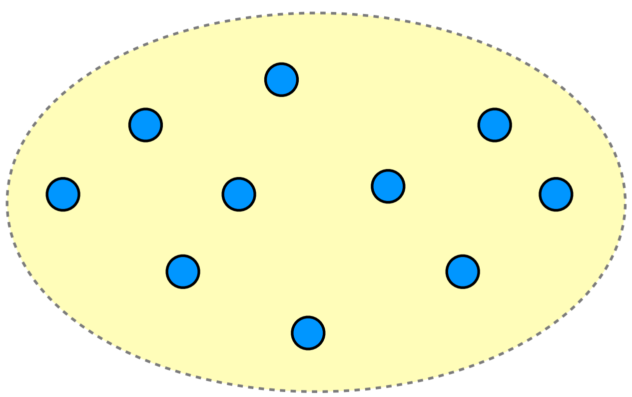
Nearest Neighbor Search
(*NNS*)

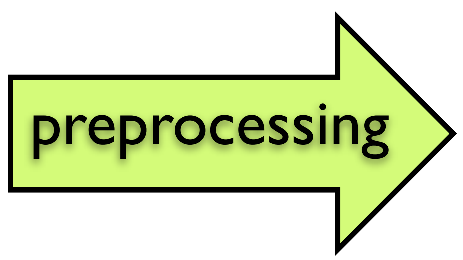
metric space (X,dist)

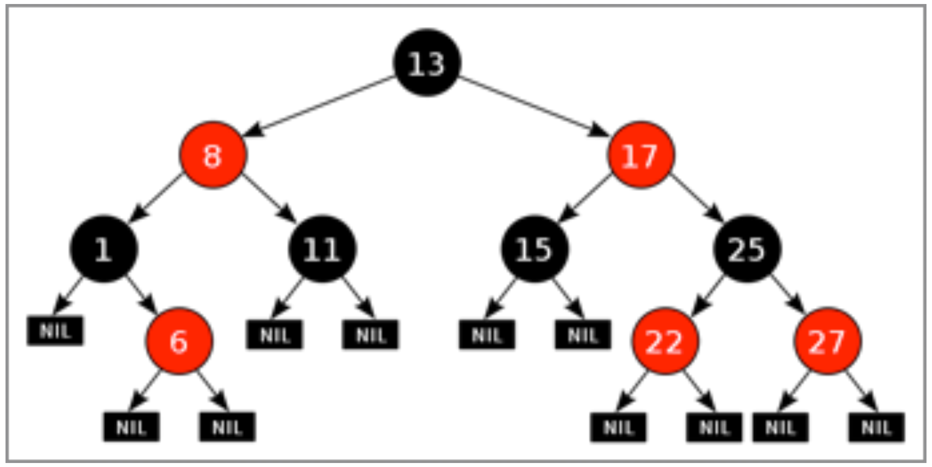
database

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$$

data structure







Nearest Neighbor Search
(*NNS*)

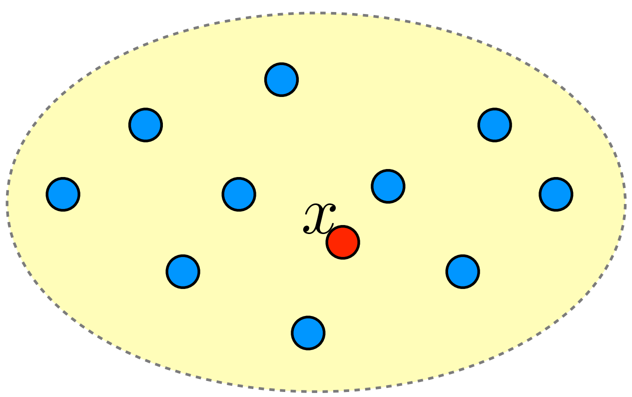
metric space (X,dist)

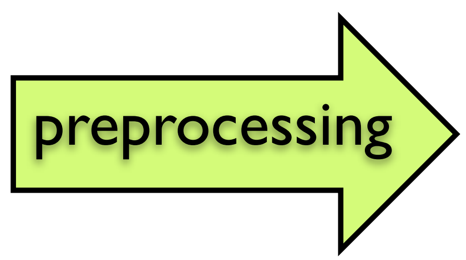
query $x \in X$

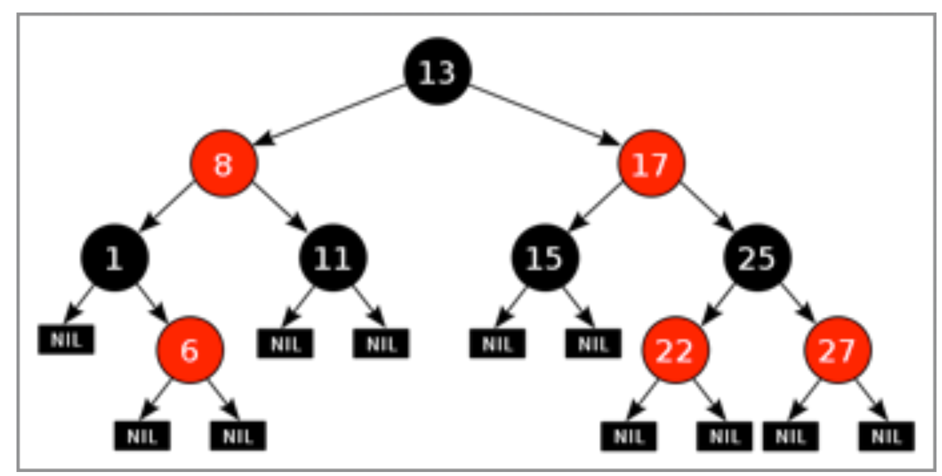
database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure







Nearest Neighbor Search
(NNS)

metric space (X, dist)

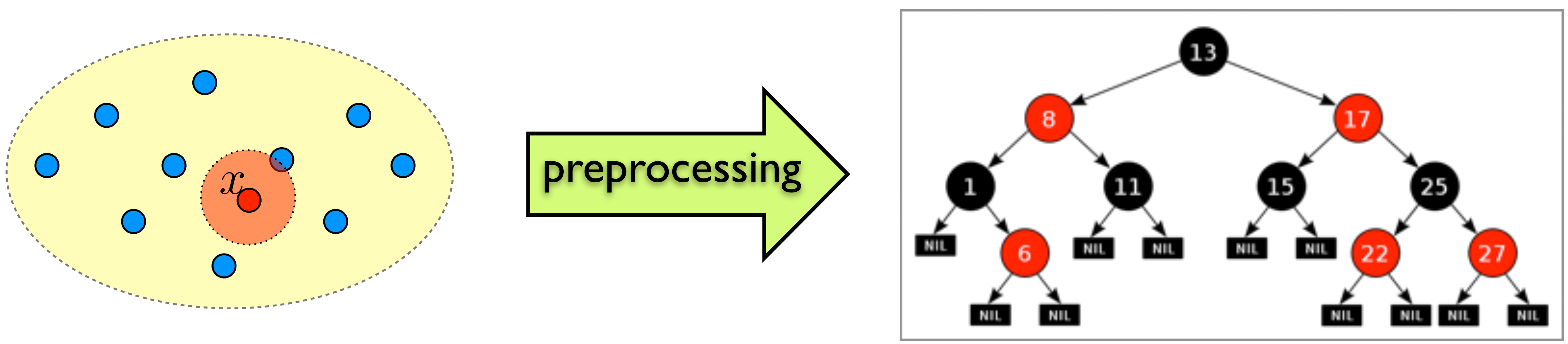
query $x \in X$

database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

access

data structure



output: database point y_i closest to the query point x

Nearest Neighbor Search
(*NNS*)

metric space (X, dist)

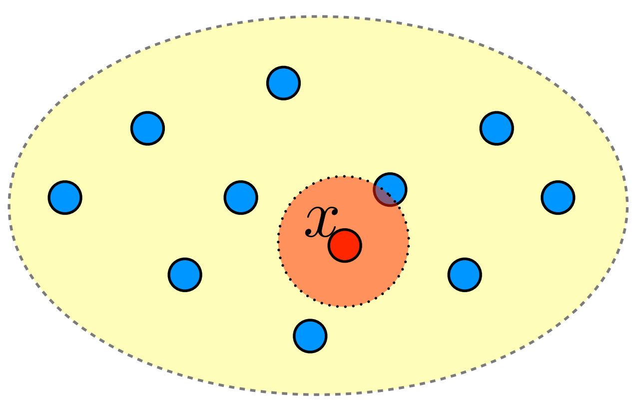
query $x \in X$

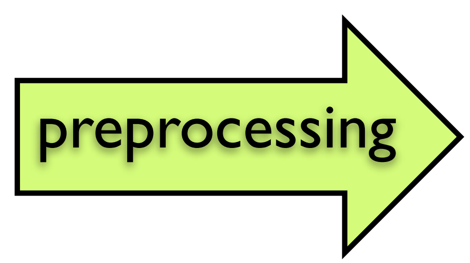
↓ access

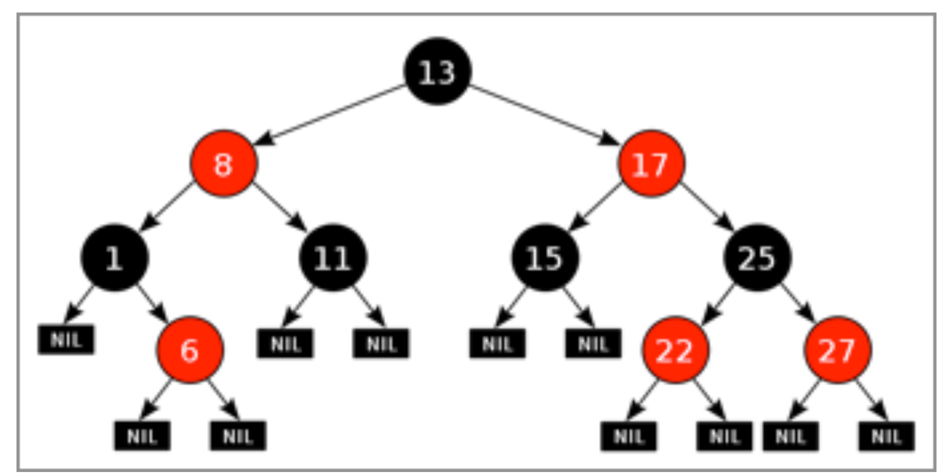
database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure







output: database point y_i closest to the query point x

applications: *database, pattern matching, machine learning, ...*

Near Neighbor Problem
(λ-*NN*)

metric space (X, dist)

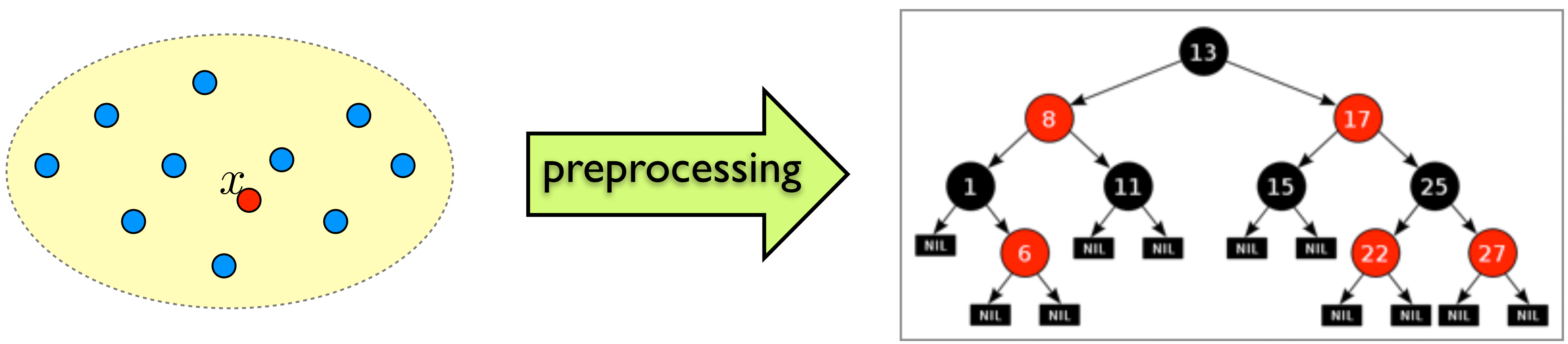
query $x \in X$

↓ access

database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure



Near Neighbor Problem

(λ-*NN*)

metric space (X, dist)

query $x \in X$

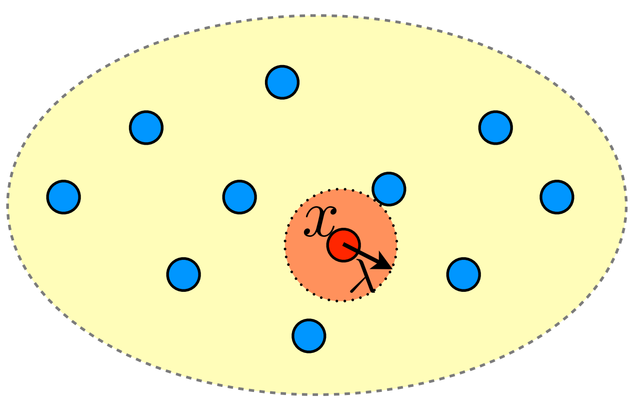
access

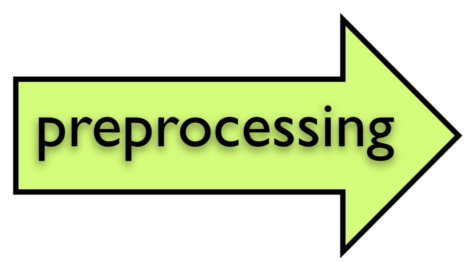
database

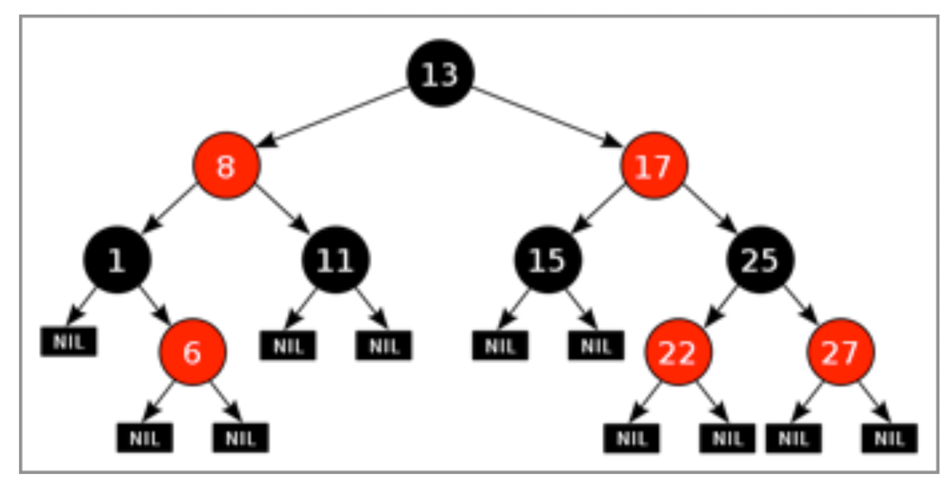
$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure

radius λ







Near Neighbor Problem

(λ-*NN*)

metric space (X,dist)

query $x \in X$

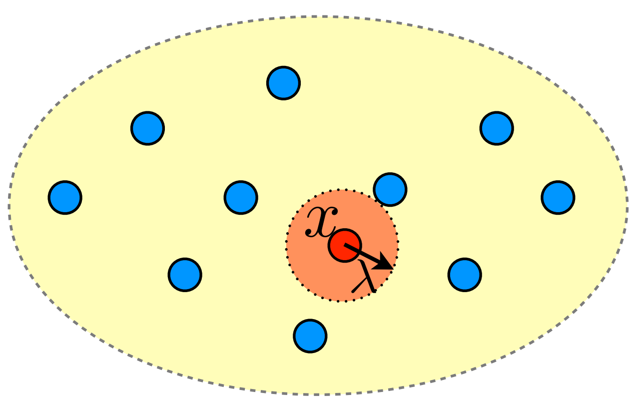
↓ access

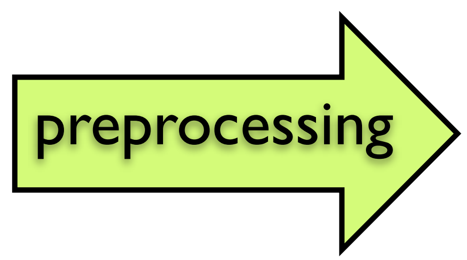
database

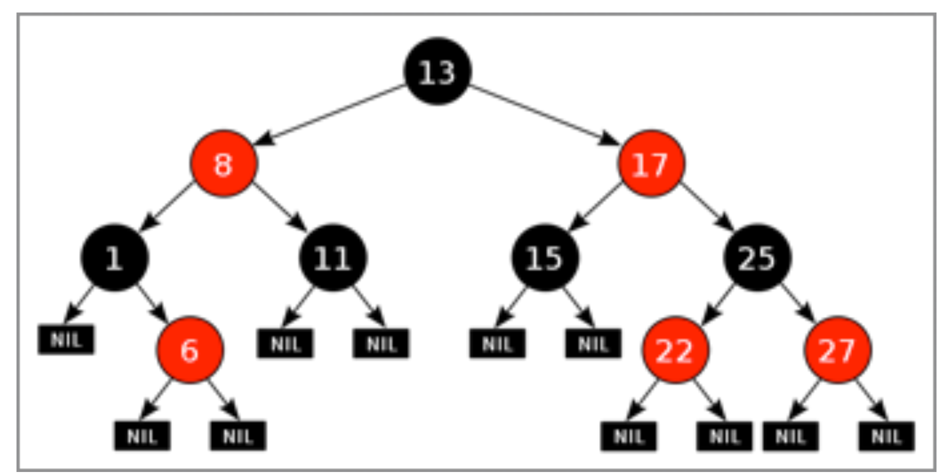
$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure

radius λ







λ-NN: answer "yes" if $\exists y_i$ that is λ-close to x

"no" if all y_i are λ-faraway from x

Approximate Near Neighbor
(ANN)

metric space (X, dist)

query $x \in X$

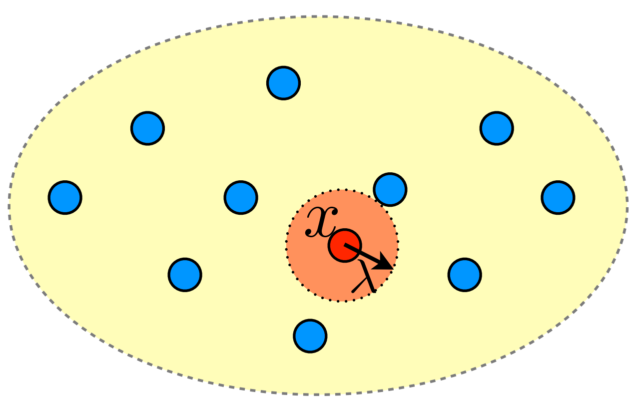
access

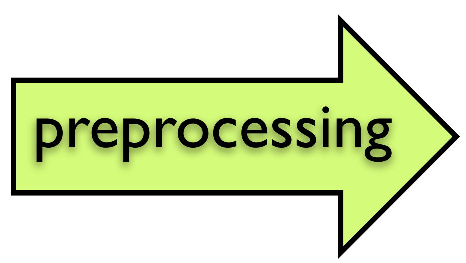
database

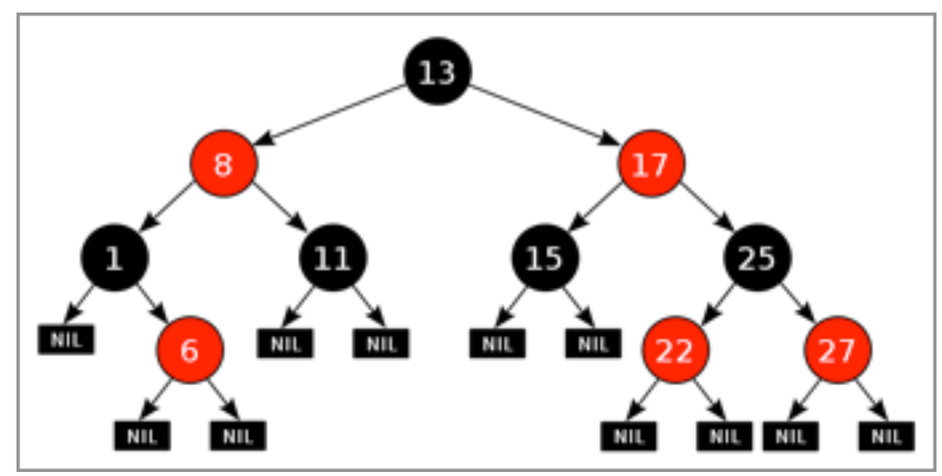
$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$$

data structure

radius λ







Approximate Near Neighbor
(*ANN*)

metric space (X, dist)

query $x \in X$

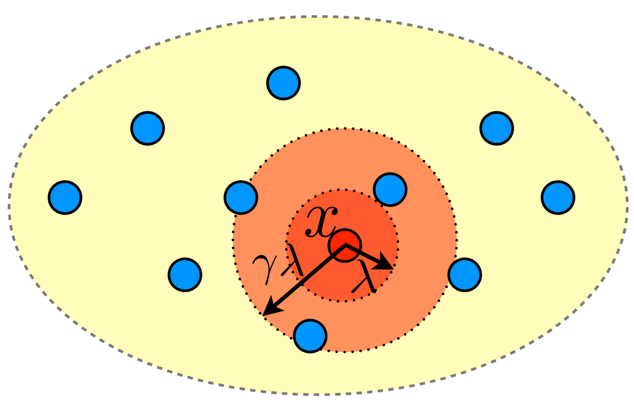
↓ access

database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

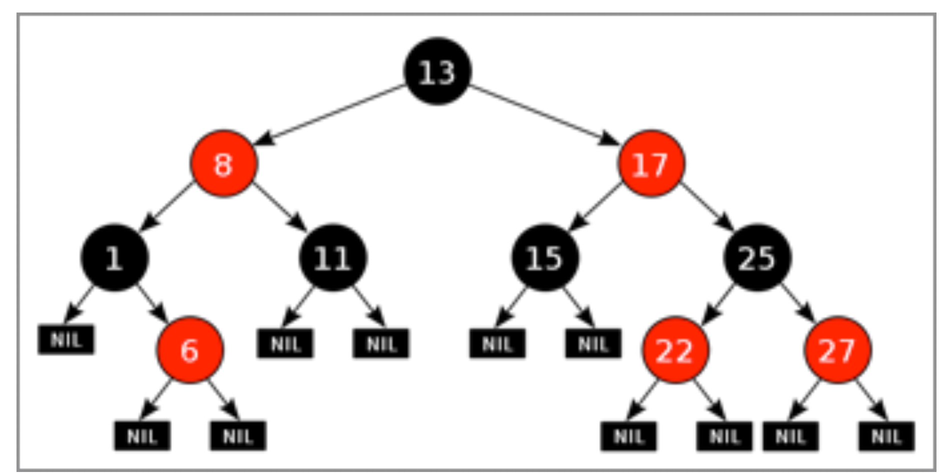
data structure

radius λ



preprocessing

approx ratio $\gamma \geq 1$



(γ, λ)-*ANN*: answer "yes" if $\exists y_i$ that is λ-close to x

"no" if all y_i are $\gamma\lambda$-faraway from x

arbitrary if otherwise

Approximate Near Neighbor
(ANN)

metric space (X,dist)

database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

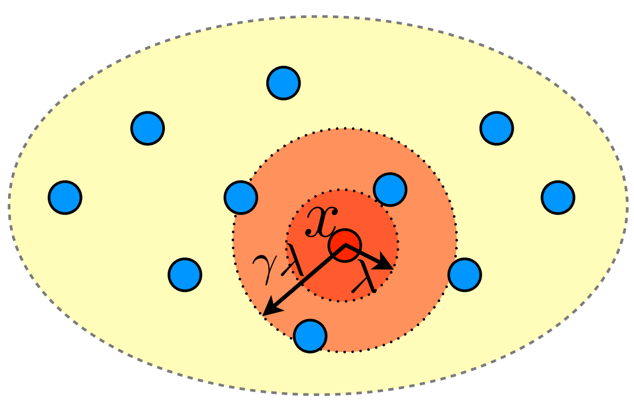
query $x \in X$

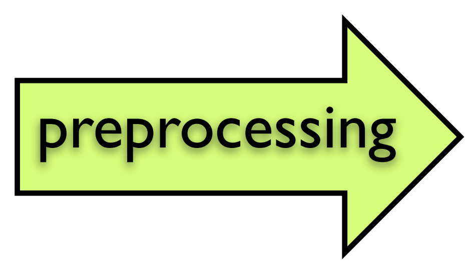
access

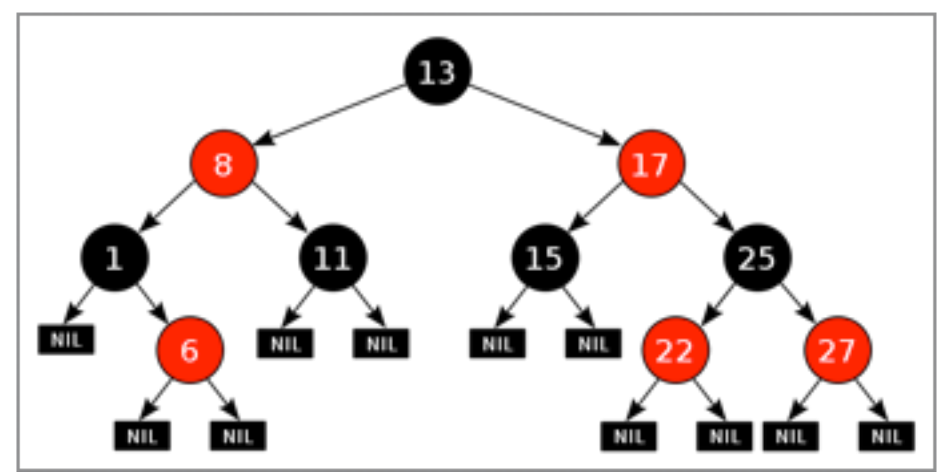
data structure

radius λ

approx ratio $\gamma \geq 1$







Hamming space $X = \{0,1\}^d$

$\text{dist}(x,z) = \|x - z\|_1$

Hamming distance

Approximate Near Neighbor
(ANN)

metric space (X, dist)

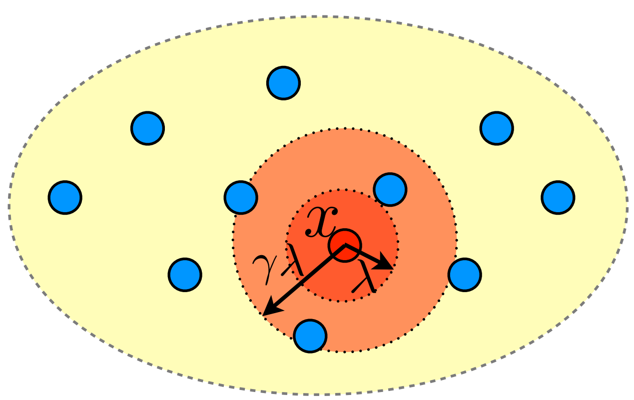
query $x \in X$

access

database

$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

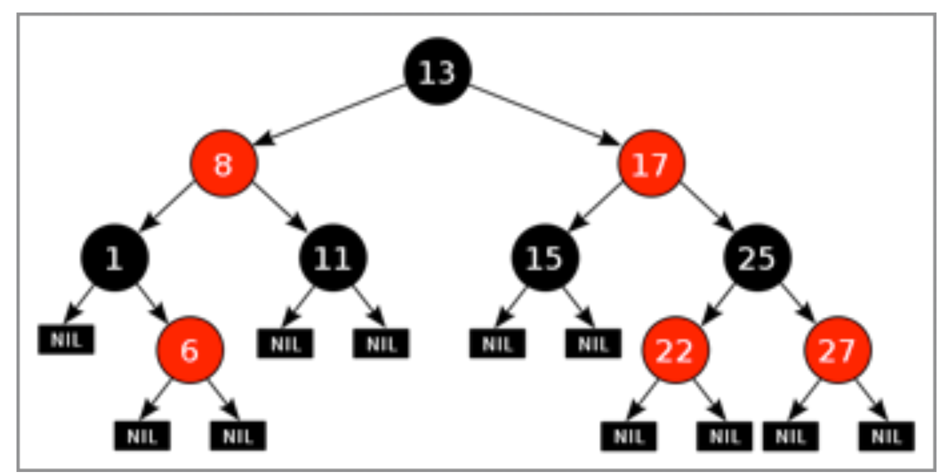
data structure



radius λ

preprocessing

approx ratio $\gamma \geq 1$



Hamming space $X = \{0,1\}^d$ $\quad$ $\text{dist}(x,z) = \|x - z\|_1$

Hamming distance

Curse of dimensionality!

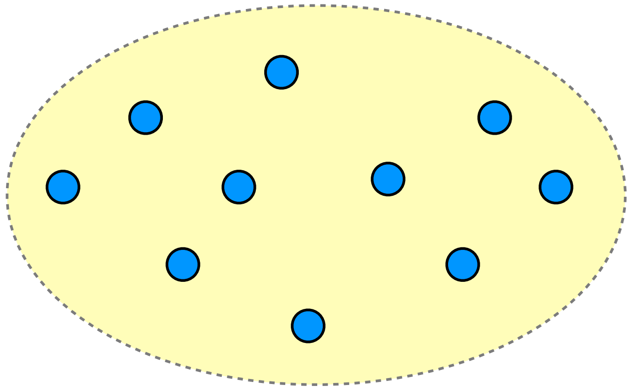
Cell-Probe Model

data structure problem:

$$f : X \times Y \to Z$$

database

$$y \in Y$$



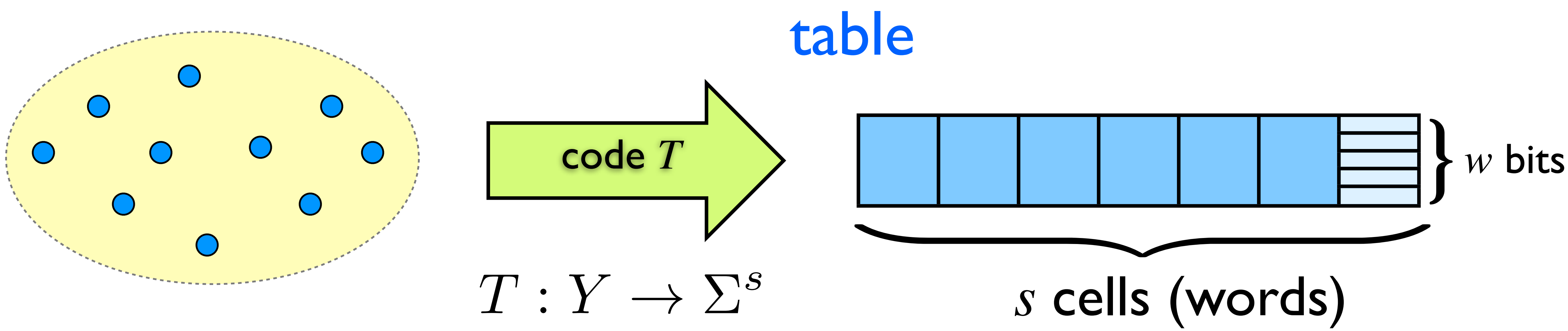
Cell-Probe Model

data structure problem:

$$f : X \times Y \to Z$$

database

$$y \in Y$$



$$T : Y \to \Sigma^s$$

where $\Sigma = \{0,1\}^w$

Cell-Probe Model

data structure problem:

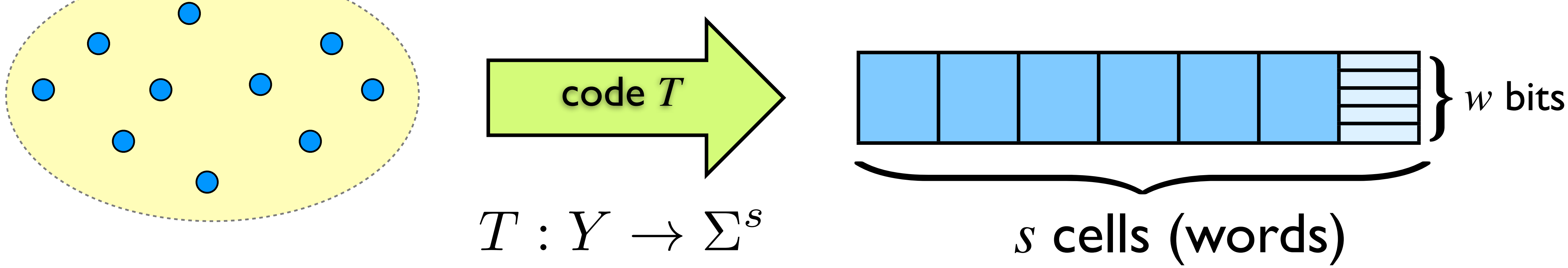
$$f : X \times Y \to Z$$

query $x \in X$

database

$$y \in Y$$

table



$$T : Y \to \Sigma^s$$

where $\Sigma = \{0,1\}^w$

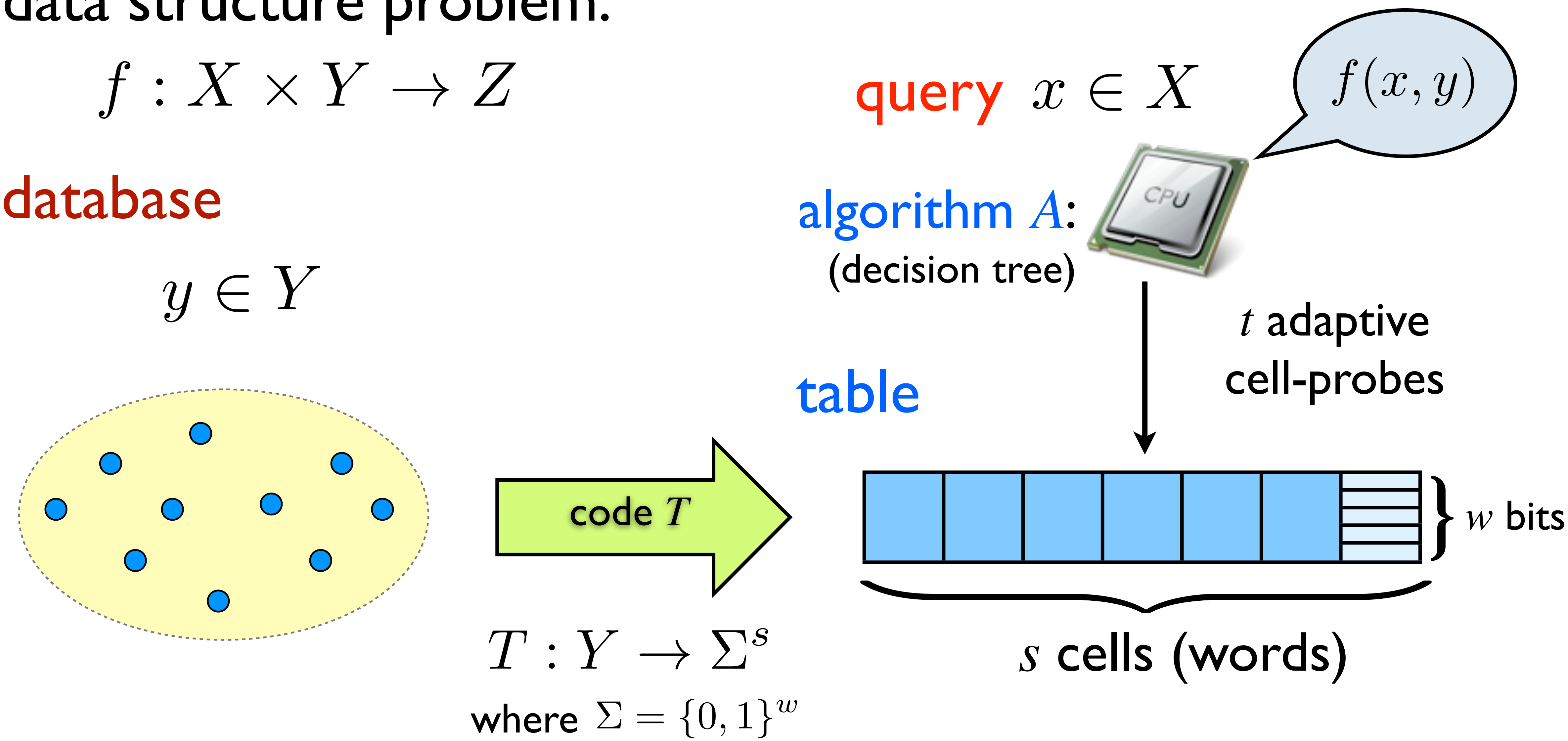
Cell-Probe Model

data structure problem:

$$f : X \times Y \rightarrow Z$$

database

$$y \in Y$$

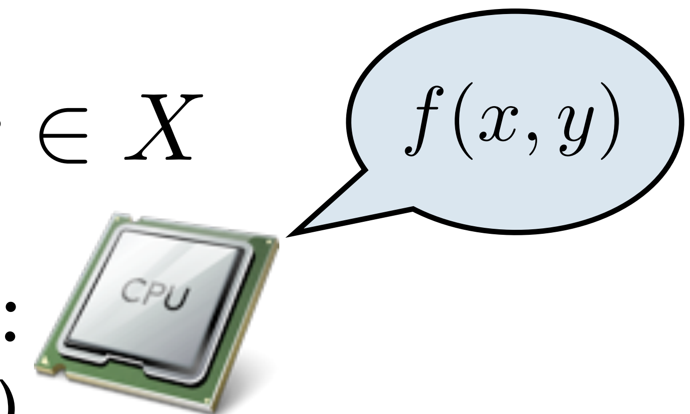


Cell-Probe Model

data structure problem:

$$f : X \times Y \to Z$$

query $x \in X$



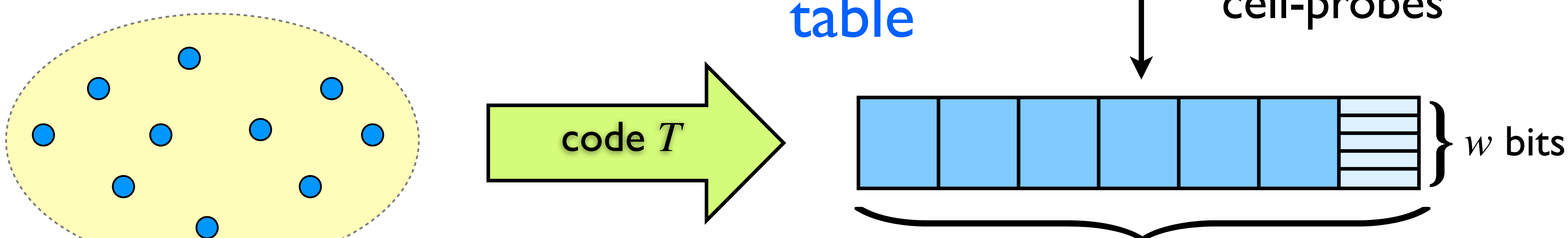
algorithm A:
(decision tree)

database

$y \in Y$

t adaptive
cell-probes

table



$T : Y \to \Sigma^s$
where $\Sigma = \{0,1\}^w$

s cells (words)

protocol: the pair (A, T)

Cell-Probe Model

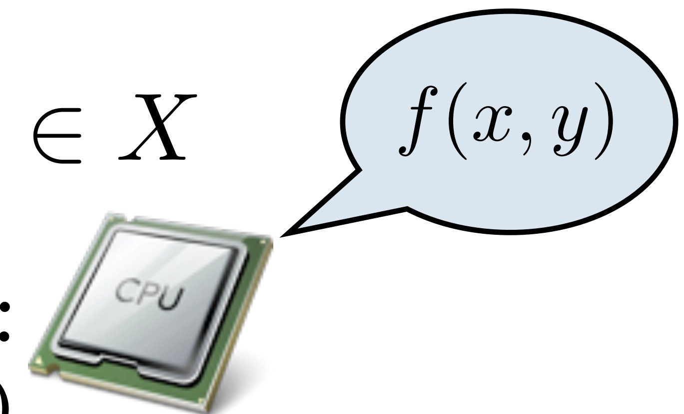
data structure problem:

$f : X \times Y \to Z$

database

$y \in Y$

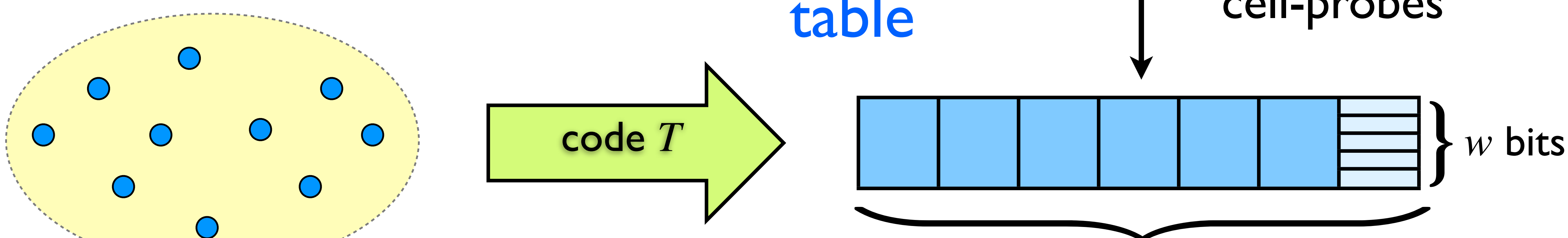
query $x \in X$



algorithm A:
(decision tree)

t adaptive
cell-probes

table



$T : Y \to \Sigma^s$
where $\Sigma = \{0,1\}^w$

s cells (words)

protocol: the pair (A, T)

(s, w, t)-cell-probing scheme

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact		
approx.		

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995]	
approx.		

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004]	

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log \log d}{\log \log \log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] for search problem $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log \log d}{\log \log \log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] by round elimination $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Average-Case Lower Bounds

data structure problem: $f : X \times Y \to Z$

(x,y) is sampled from a distribution *D* over $X \times Y$

Average-Case Lower Bounds

data structure problem: $f : X \times Y \to Z$

(x,y) is sampled from a distribution D over $X \times Y$

- deterministic or *Les Vegas* randomized algorithm: $f(x,y)$ is returned in $t(x,y)$ cell-probes
  - $\mathbf{E}_D[t(x, y)] \le t$

Average-Case Lower Bounds

data structure problem: $f : X \times Y \to Z$

(x,y) is sampled from a distribution *D* over $X \times Y$

- deterministic or *Les Vegas* randomized algorithm: $f(x,y)$ is returned in $t(x,y)$ cell-probes
  - $\mathbf{E}_D[t(x, y)] \leq t$
- *Monte Carlo* randomized algorithm:
  - $\Pr[\, f(x,y)$ is correctly returned in t cell-probes$] > 2/3$

Average-Case Lower Bounds

data structure problem: $f : X \times Y \to Z$

(x,y) is sampled from a distribution *D* over $X \times Y$

- deterministic or *Les Vegas* randomized algorithm:
  $f(x,y)$ is returned in $t(x,y)$ cell-probes
  - $\mathbf{E}_D[t(x, y)] \leq t$
- *Monte Carlo* randomized algorithm:
  - Pr[$f(x,y)$ is correctly returned in t cell-probes] > 2/3
- In data-dependent LSH [Andoni Razenshteyn 2015]:
  a key step is to solve the problem on random input.

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ average-case [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] average-case $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ average-case [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] average-case $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] worst-case $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ average-case [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] average-case $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] worst-case
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] worst-case $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ average-case [Panigrahy Talwar Wieder 2008, 2010]

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] average-case $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] worst-case
approx.	$t = \Omega\left(\frac{d}{\log s}\right)$ [Liu 2004] $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006] $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]	$t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] worst-case $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ average-case [Panigrahy Talwar Wieder 2008, 2010]

worst-case

Approximate Near Neighbor

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in Hamming space $\{0,1\}^d$

Approximate Near Neighbor

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in Hamming space $\{0,1\}^d$

that solves the problem with t cell-probes *in expectation*
on a table of $s<2^d$ cells, each of $w<n^{o(1)}$ bits,
under the *hard distribution*:

Approximate Near Neighbor

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in Hamming space $\{0,1\}^d$

that solves the problem with t cell-probes *in expectation*
on a table of $s<2^d$ cells, each of $w<n^{o(1)}$ bits,
under the *hard distribution*:

database: $y=(y_1, y_2,\ldots,y_n)$ with $y_i\in\{0,1\}^d$ *i.i.d.* uniform
query: uniform and independent $x\in\{0,1\}^d$

Approximate Near Neighbor

deterministic or *LV randomized* algorithm for

(γ,λ)-***ANN*** in Hamming space $\{0,1\}^d$

that solves the problem with t cell-probes *in expectation* on a table of $s<2^d$ cells, each of $w<n^{o(1)}$ bits, under the *hard distribution*:

database: $y=(y_1, y_2,\ldots,y_n)$ with $y_i\in\{0,1\}^d$ *i.i.d.* uniform

query: uniform and independent $x\in\{0,1\}^d$

$$t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$$

Metric Expansion

metric space (X,dist)

λ-neighborhood: $\forall x \in X$, denote $N_\lambda(x) = \{z \in X \mid \text{dist}(x,z) \le \lambda\}$

$\forall A \subseteq X$, denote $N_\lambda(A) = \{z \in X \mid \exists x \in A \text{ s.t. } \text{dist}(x,z) \le \lambda\}$

Metric Expansion

metric space (X,dist)

λ-neighborhood: $\forall x \in X$, denote $N_\lambda(x) = \{z \in X \mid \mathrm{dist}(x,z) \leq \lambda\}$

$\forall A \subseteq X$, denote $N_\lambda(A) = \{z \in X \mid \exists x \in A \text{ s.t. } \mathrm{dist}(x,z) \leq \lambda\}$

μ: probability distribution over X

Metric Expansion

metric space (X,dist)

λ-neighborhood: $\forall x \in X$, denote $N_\lambda(x) = \{z \in X \mid \text{dist}(x,z) \leq \lambda\}$

$\forall A \subseteq X$, denote $N_\lambda(A) = \{z \in X \mid \exists x \in A \text{ s.t. } \text{dist}(x,z) \leq \lambda\}$

μ: probability distribution over X

In a metric space (X,dist), we say:

- λ-neighborhoods are weakly independent under μ:

$$\forall x \in X,\ \mu(N_\lambda(x)) < 0.99/n$$

Metric Expansion

metric space (X,dist)

λ-neighborhood: $\forall x \in X$, denote $N_\lambda(x) = \{z \in X \mid \text{dist}(x,z) \le \lambda\}$

$\forall A \subseteq X$, denote $N_\lambda(A) = \{z \in X \mid \exists x \in A \text{ s.t. } \text{dist}(x,z) \le \lambda\}$

μ: probability distribution over X

In a metric space (X,dist), we say:

- λ-neighborhoods are weakly independent under μ:

$$\forall x \in X,\ \ \mu(N_\lambda(x)) < 0.99/n$$

- λ-neighborhoods are (Φ,Ψ)-expanding under μ:

$$\forall A \subseteq X,\ \ \mu(A) \ge 1/\Phi \Rightarrow \mu(N_\lambda(A)) \ge 1\text{-}1/\Psi$$

Metric Expansion

metric space (X,dist)

λ-neighborhood: $\forall x \in X$, denote $N_\lambda(x) = \{z \in X \mid \text{dist}(x,z) \le \lambda\}$

$\forall A \subseteq X$, denote $N_\lambda(A) = \{z \in X \mid \exists x \in A \text{ s.t. } \text{dist}(x,z) \le \lambda\}$

μ: probability distribution over X

In a metric space (X,dist), we say:

- λ-neighborhoods are weakly independent under μ:

$$\forall x \in X,\ \ \mu(N_\lambda(x)) < 0.99/n$$

- λ-neighborhoods are (Φ,Ψ)-expanding under μ:

$$\forall A \subseteq X,\ \ \mu(A) \ge 1/\Phi \Rightarrow \mu(N_\lambda(A)) \ge 1\text{-}1/\Psi$$

vertex expansion, “blow-up” effect

for metric space (X, dist), distribution μ over X:

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in metric space (X, dist)

for metric space (X,dist), distribution μ over X:

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in metric space (X,dist)

that solves the problem with t cell-probes *in expectation* on a table of size (s, w), under the *hard distribution* $\mu\times\mu^n$:

query $x \sim \mu$, database $y=(y_1, y_2,\ldots,y_n) \sim \mu^n$

for metric space (X,dist), distribution μ over X:

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

$$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in metric space (X,dist)

that solves the problem with t cell-probes *in expectation* on a table of size (s, w), under the *hard distribution* $\mu\times\mu^n$:

query $x \sim \mu$, database $y=(y_1, y_2,\ldots,y_n) \sim \mu^n$

for metric space (X,dist), distribution μ over X:

- $\gamma\lambda$-neighborhoods are weakly independent under μ:
  $$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$
- λ-neighborhoods are (Φ,Ψ)-expanding under μ:
  $$\forall A \subseteq X \text{ that } \mu(A) \geq 1/\Phi \Rightarrow \mu(N_\lambda(A)) \geq 1\text{-}1/\Psi$$

deterministic or *LV randomized* algorithm for

(γ, λ)-***ANN*** in metric space (X,dist)

that solves the problem with t cell-probes *in expectation* on a table of size (s, w), under the *hard distribution* $\mu\times\mu^n$:

query $x \sim \mu$, database $y=(y_1, y_2,\ldots,y_n) \sim \mu^n$

for metric space (X,dist), distribution μ over X:

- $\gamma\lambda$-neighborhoods are weakly independent under μ:
  $$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$
- λ-neighborhoods are (Φ,Ψ)-expanding under μ:
  $$\forall A \subseteq X \text{ that } \mu(A) \geq 1/\Phi \Rightarrow \mu(N_\lambda(A)) \geq 1\text{-}1/\Psi$$

deterministic or *LV randomized* algorithm for

(γ,λ)-***ANN*** in metric space (X,dist)

that solves the problem with t cell-probes *in expectation* on a table of size (s, w), under the *hard distribution* $\mu\times\mu^n$:

query $x \sim \mu$, database $y=(y_1, y_2,\ldots,y_n) \sim \mu^n$

$$\Rightarrow \quad t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right) \quad \text{or} \quad t = \Omega\left(\frac{n \log \Psi}{w + \log s}\right)$$

Asymmetric Communications

$$f : X \times Y \to \{0, 1\}$$

$x \in X$   $y \in Y$

Asymmetric Communications

$f(x, y)$

$$f : X \times Y \to \{0, 1\}$$

$x \in X$

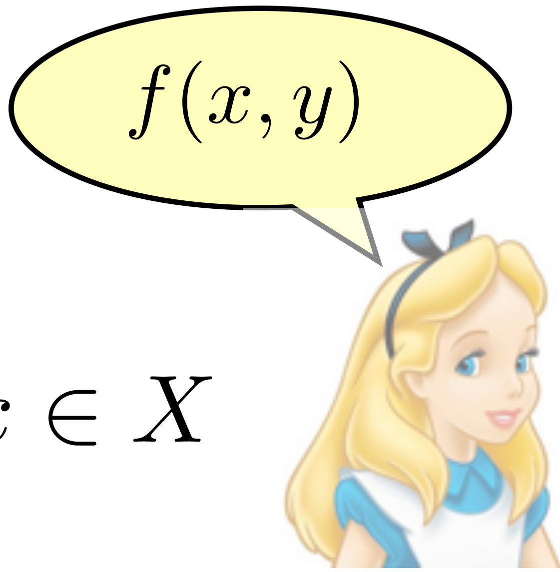
a bits in total →

← b bits in total

$y \in Y$

[a,b]-protocol: Alice sends a total of ≤a bits
Bob sends a total of ≤b bits

Asymmetric Communications



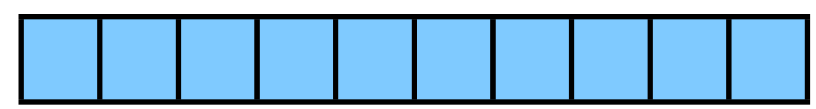
$f : X \times Y \to \{0,1\}$

$x \in X$

a bits in total →

← b bits in total

$y \in Y$



[a,b]-protocol: Alice sends a total of ≤a bits
Bob sends a total of ≤b bits

(s,w,t)-cell-probing scheme ⇒ [$t \log s$, tw]-protocol

The *Richness* Lemma

$$f : X \times Y \to \{0,1\}$$

[a,b]-protocol: Alice sends a total of $\leq a$ bits
Bob sends a total of $\leq b$ bits

The *Richness* Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

[a,b]-protocol: Alice sends a total of ≤a bits
Bob sends a total of ≤b bits

The *Richness* Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

α-dense: density of 1s $\geq \alpha$ under $\mu \times \nu$

$[a,b]$-protocol: Alice sends a total of $\leq a$ bits
Bob sends a total of $\leq b$ bits

The *Richness* Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

α-dense: density of 1s $\geq \alpha$ under $\mu \times \nu$

monochromatic 1-rectangle: $A \times B$ with $A \subseteq X$, $B \subseteq Y$
$\forall (x,y) \in A \times B$, $f(x,y)=1$

$[a,b]$-protocol: Alice sends a total of $\leq a$ bits
Bob sends a total of $\leq b$ bits

The *Richness* Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

α-dense: density of 1s $\geq \alpha$ under $\mu\times\nu$

monochromatic 1-rectangle: $A\times B$ with $A\subseteq X$, $B\subseteq Y$

$\forall (x,y)\in A\times B,\ f(x,y)=1$

$[a,b]$-protocol: Alice sends a total of $\leq a$ bits

Bob sends a total of $\leq b$ bits

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu\times\nu$

f has $[a,b]$-protocol

$\Rightarrow$ f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(a)} \\ \nu(B) \geq 2^{-O(a+b)} \end{cases}$$

The *Richness* Lemma

$f : X \times Y \to \{0, 1\}$ distributions μ over X, ν over Y

α-dense: density of 1s $\geq \alpha$ under $\mu \times \nu$

$[a,b]$-protocol: Alice sends a total of $\leq a$ bits
Bob sends a total of $\leq b$ bits

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu \times \nu$
f has $[a,b]$-protocol
⇒
f has 1-rectangle $A \times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(a)} \\ \nu(B) \geq 2^{-O(a+b)} \end{cases}$$

The *Richness* Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

α-dense: density of 1s $\geq \alpha$ under $\mu \times \nu$

$[a,b]$-protocol: Alice sends a total of $\leq a$ bits
Bob sends a total of $\leq b$ bits

(s,w,t)-cell-probing scheme ⇒ $[t \log s, tw]$-protocol

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu \times \nu$ and f has $[a,b]$-protocol ⇒ f has 1-rectangle $A \times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(a)} \\ \nu(B) \geq 2^{-O(a+b)} \end{cases}$$

The *Richness* Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

α-dense: density of 1s $\geq \alpha$ under $\mu \times \nu$

$[a,b]$-protocol: Alice sends a total of $\leq a$ bits
Bob sends a total of $\leq b$ bits

(s,w,t)-cell-probing scheme ⇒ $[t \log s, tw]$-protocol

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu \times \nu$
f has (s,w,t)-cell-probing scheme

⇒ f has 1-rectangle $A \times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log s)} \\ \nu(B) \geq 2^{-O(t \log s + tw)} \end{cases}$$

A New Richness Lemma

$f : X \times Y \to \{0, 1\}$ distributions μ over X, ν over Y

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu \times \nu$
f has (s,w,t)-cell-probing scheme

⇒ f has 1-rectangle $A \times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log s)} \\ \nu(B) \geq 2^{-O(t \log s + tw)} \end{cases}$$

A New Richness Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu\times\nu$
f has (s,w,t)-cell-probing scheme

⇒ f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log s)} \\ \nu(B) \geq 2^{-O(t \log s + tw)} \end{cases}$$

New Richness lemma

f is 0.01-dense under $\mu\times\nu$
f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

⇒ $\forall \Delta \in [320000t, s]$,
f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

A New Richness Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

Richness lemma (Miltersen, Nisan, Safra, Wigderson, 1995)

f is 0.01-dense under $\mu\times\nu$
f has (s,w,t)-cell-probing scheme

⇒ f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log s)} \\ \nu(B) \geq 2^{-O(t \log s + tw)} \end{cases}$$

New Richness lemma

f is 0.01-dense under $\mu\times\nu$
f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

⇒ $\forall \Delta \in [320000t, s]$,
f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

when $\Delta = O(t)$, it becomes the richness lemma (with slightly better bounds)

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

⇒

$\forall\Delta\in[320000t, s]$,

f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

$\Rightarrow$ $\forall\Delta\in[320000t, s]$,

f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

⟹

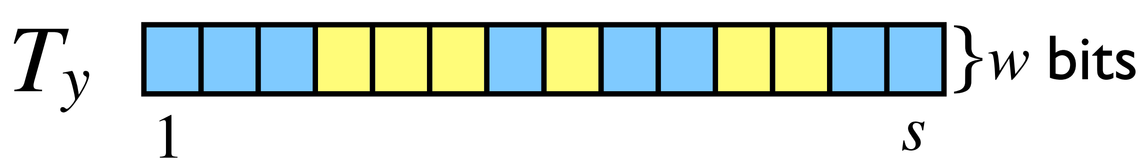
$\forall\Delta\in[320000t, s]$,

f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

∃ constant fraction (under ν) of "good" databases y:

∀good y,



New Richness lemma

f is 0.01-dense under $\mu \times \nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu \times \nu$

$\Rightarrow$

$\forall \Delta \in [320000t, s]$,

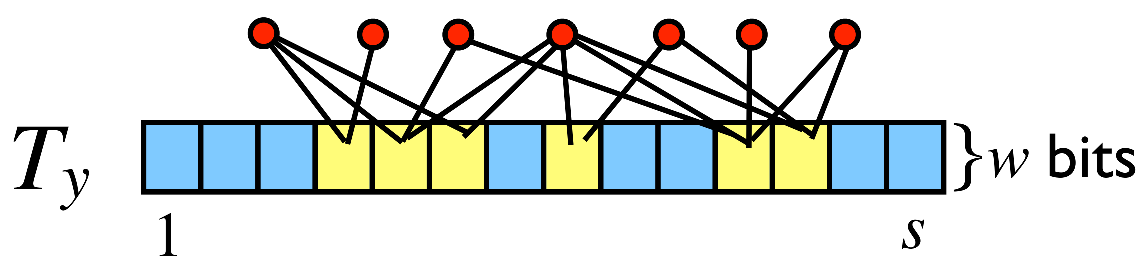
f has 1-rectangle $A \times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists$ Δ cells resolving $2^{-O(t \log (s/\Delta))}$ fraction (under μ) positive queries



New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

$\Rightarrow$

$\forall\Delta\in[320000t, s]$,

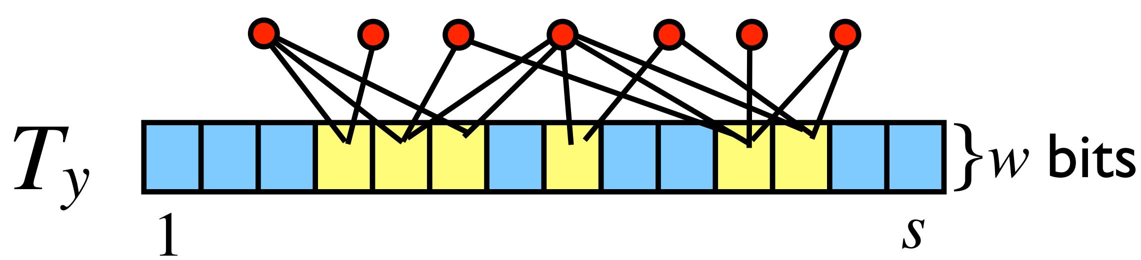
f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists\ \Delta$ cells resolving $2^{-O(t\log(s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

$\Rightarrow$ $\forall\Delta\in[320000t, s]$,

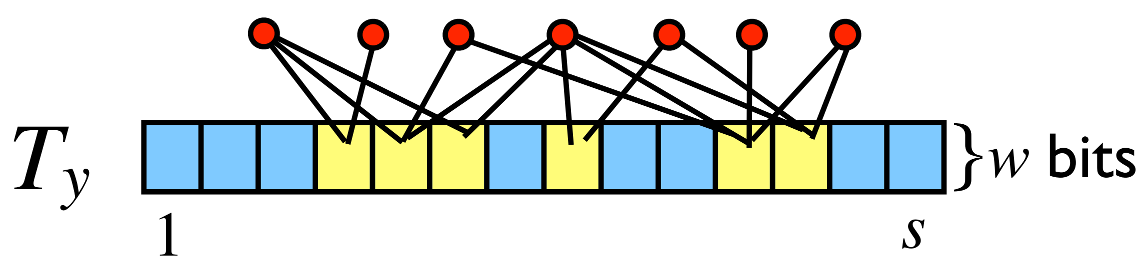
f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \ge 2^{-O(t\log(s/\Delta))} \\ \nu(B) \ge 2^{-O(\Delta\log(s/\Delta)+\Delta w)} \end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists$ Δ cells resolving $2^{-O(t\log(s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

good $y \longmapsto \omega$

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

$\Rightarrow$

$\forall\Delta\in[320000t, s]$,

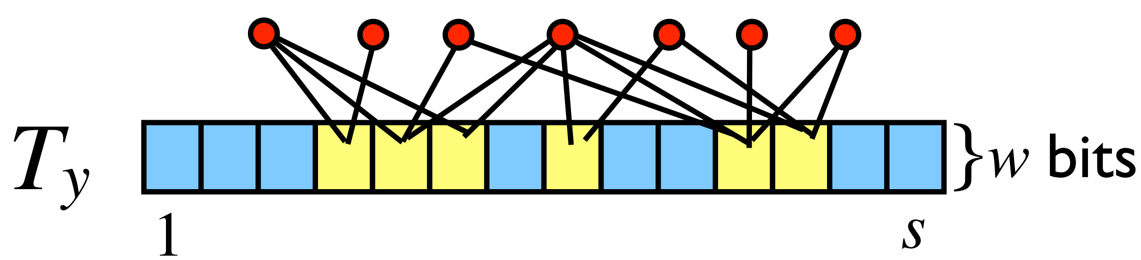
f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists$ Δ cells resolving $2^{-O(t \log (s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

good $y \longmapsto \omega$

$\leq \binom{s}{\Delta} 2^{\Delta w} = 2^{O(\Delta \log \frac{s}{\Delta} + \Delta w)}$ possibilities

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

$\Rightarrow$ $\forall \Delta\in[320000t, s]$,

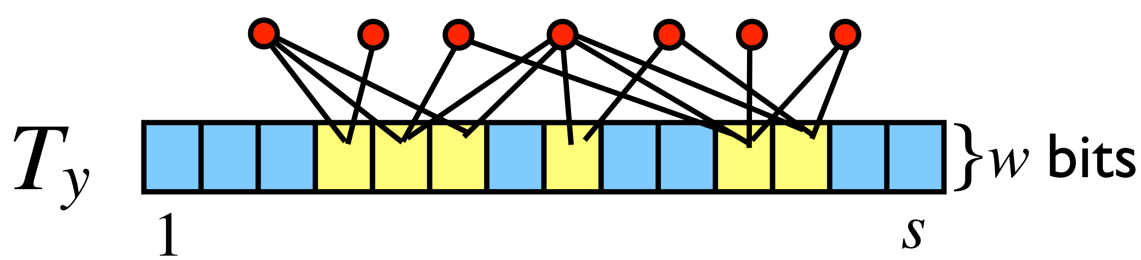
f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\ge 2^{-O(t\log(s/\Delta))}\\ \nu(B)\ge 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists$ Δ cells resolving $2^{-O(t\log(s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

good $y\longmapsto\omega$ $\quad\le\binom{s}{\Delta}2^{\Delta w}=2^{O(\Delta\log\frac{s}{\Delta}+\Delta w)}$ possibilities

$\ge 2^{-O(\Delta\log(s/\Delta)+\Delta w)}$ fraction (under ν) good $y\longmapsto$ the same ω

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

$\Rightarrow$

$\forall\Delta\in[320000t, s]$,

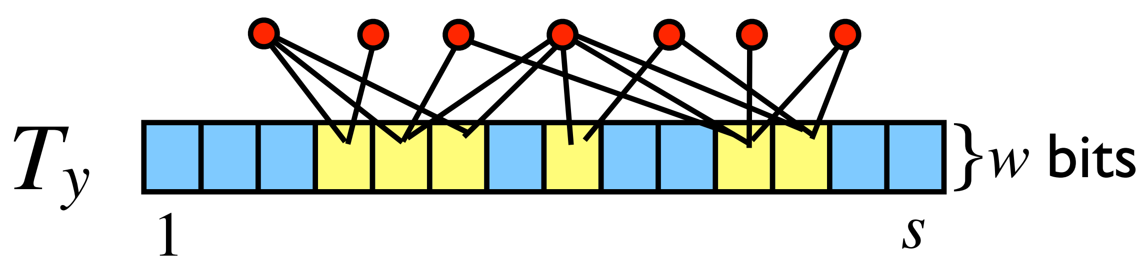
f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists$ Δ cells resolving $2^{-O(t\log(s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

good $y \longmapsto \omega$

$\leq \binom{s}{\Delta}2^{\Delta w} = 2^{O(\Delta\log\frac{s}{\Delta}+\Delta w)}$ possibilities

$\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}$ fraction (under ν) good $y \longmapsto$ the same ω

cell-probe model: once ω is fixed,

the set of positive queries resolved by ω is fixed

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

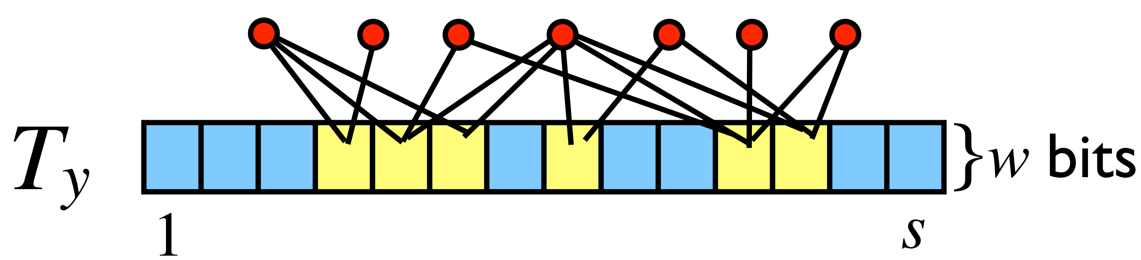
$\Rightarrow$ $\forall\Delta\in[320000t, s]$, f has 1-rectangle $A\times B$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists\ \Delta$ cells resolving $2^{-O(t\log(s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

good $y\longmapsto\omega$ — $\leq\binom{s}{\Delta}2^{\Delta w}=2^{O(\Delta\log\frac{s}{\Delta}+\Delta w)}$ possibilities

B : $\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}$ fraction (under ν) good $y\longmapsto$ the same ω

cell-probe model: once ω is fixed,
the set of positive queries resolved by ω is fixed

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

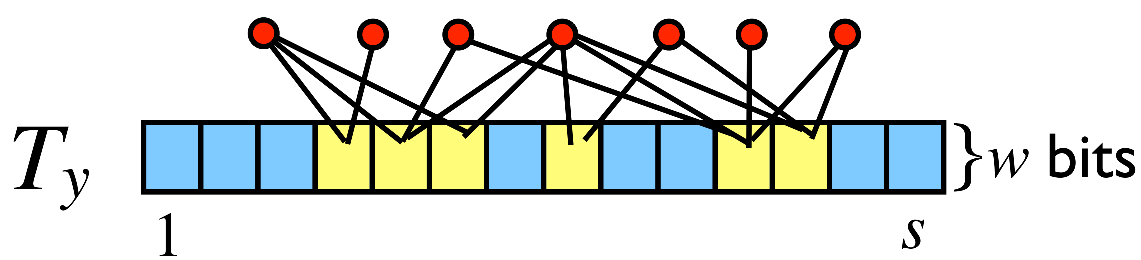
$\Rightarrow$ $\forall \Delta \in [320000t, s]$, f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

$\exists$ constant fraction (under ν) of "good" databases y:

$\forall$good y,

$\exists$ Δ cells resolving $2^{-O(t \log (s/\Delta))}$ fraction (under μ) positive queries



ω: positions & contents of these Δ cells

good $y \longmapsto \omega$ $\quad \leq \binom{s}{\Delta} 2^{\Delta w} = 2^{O(\Delta \log \frac{s}{\Delta} + \Delta w)}$ possibilities

B : $\geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)}$ fraction (under ν) good $y \longmapsto$ the same ω

cell-probe model: once ω is fixed,

A : the set of positive queries resolved by ω is fixed

$f : X \times Y \to \{0, 1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu\times\nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \leq t$

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu\times\nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \le t$

fix database (column) $y\in Y$

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu\times\nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \le t$

fix database (column) $y\in Y$

set of positive queries: $X_y^+ = \{x \in X \mid f(x,y) = 1\}$

$f : X \times Y \to \{0, 1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu \times \nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \le t$

fix database (column) $y \in Y$

set of positive queries: $X_y^+ = \{x \in X \mid f(x, y) = 1\}$

distribution over positive queries induced by μ:

$$\mu_y^+(x) = \mu(x \mid X_y^+) = \Pr_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} = x \mid f(\boldsymbol{x}, y) = 1]$$

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu\times\nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \le t$

fix database (column) $y\in Y$

set of positive queries: $X_y^+ = \{x \in X \mid f(x,y) = 1\}$

distribution over positive queries induced by μ:

$$\mu_y^+(x) = \mu(x \mid X_y^+) = \Pr_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} = x \mid f(\boldsymbol{x}, y) = 1]$$

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \ge 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \ge 0.005$;
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \le 80000t$.

$f : X \times Y \to \{0, 1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu\times\nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \le t$

fix database (column) $y\in Y$

set of positive queries: $X_y^+ = \{x \in X \mid f(x, y) = 1\}$

distribution over positive queries induced by μ:

$$\mu_y^+(x) = \mu(x \mid X_y^+) = \Pr_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} = x \mid f(\boldsymbol{x}, y) = 1]$$

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \ge 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \ge 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \le 80000t$.

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

f is 0.01-dense under $\mu\times\nu$

$f(x,y)$ is answered with $t(x,y)$ cell-probes: $\mathbf{E}_{\mu\times\nu}[t(x,y)] \le t$

fix database (column) $y\in Y$

set of positive queries: $X_y^+ = \{x \in X \mid f(x,y) = 1\}$

distribution over positive queries induced by μ:

$$\mu_y^+(x) = \mu(x \mid X_y^+) = \Pr_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} = x \mid f(\boldsymbol{x}, y) = 1]$$

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \ge 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \ge 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \le 80000t$. (bounded average cost over positive queries)

$\exists Y_{\mathsf{good}} \subseteq Y$ s.t. $\nu(Y_{\mathsf{good}}) \geq 0.0025$ and $\forall y \in Y_{\mathsf{good}}$:

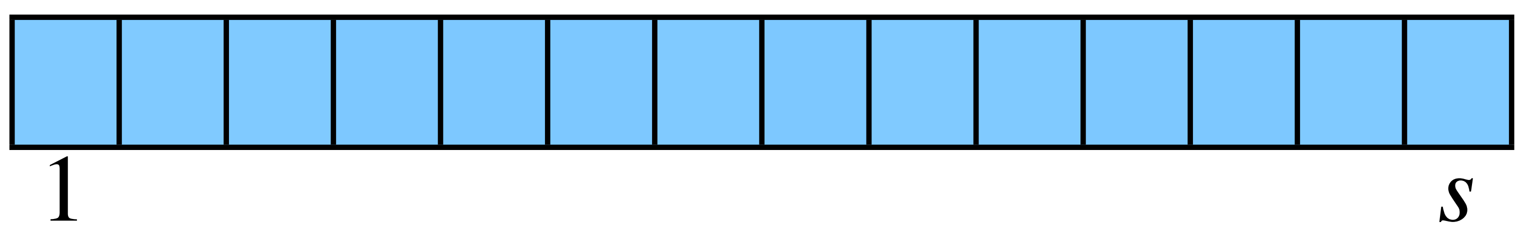
- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x} \sim \mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

$\exists Y_{\mathsf{good}} \subseteq Y$ s.t. $\nu(Y_{\mathsf{good}}) \geq 0.0025$ and $\forall y \in Y_{\mathsf{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x} \sim \mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t.$ (bounded average cost over positive queries)

fix $\forall y \in Y_{\mathsf{good}}$:

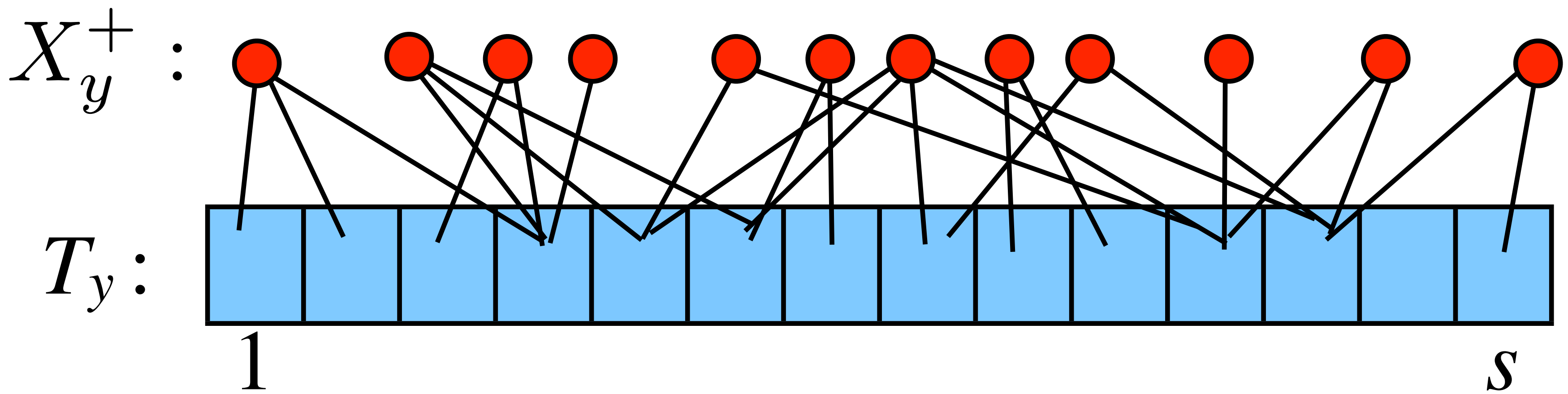
T_y:



$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x} \sim \mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

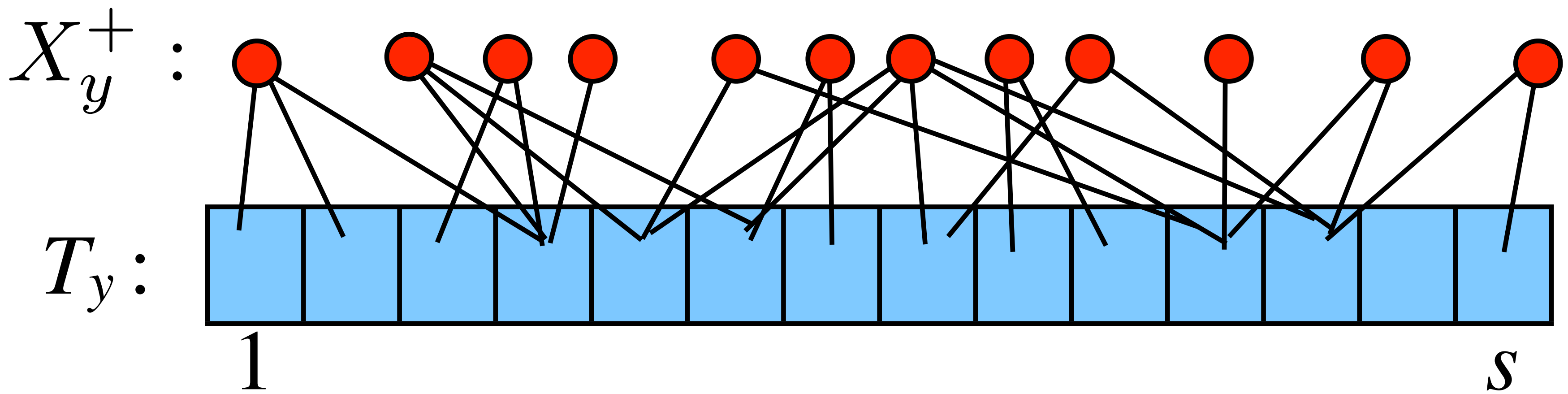
fix $\forall y \in Y_{\text{good}}$:



$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x} \sim \mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

fix $\forall y \in Y_{\text{good}}$:

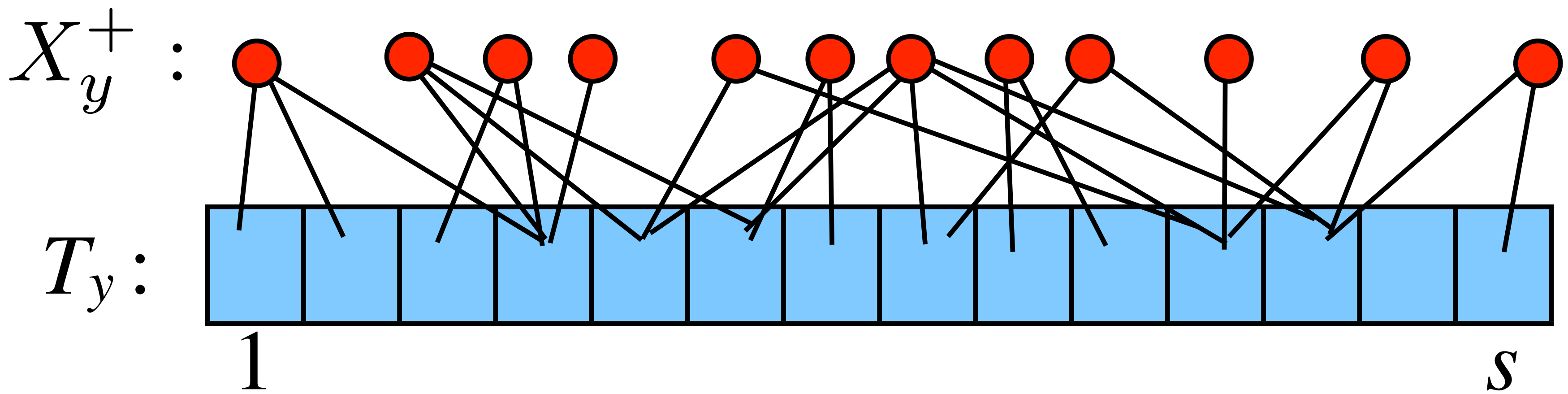


hypergraph with vertices $[s]$ and hyperedges X_y^+

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

fix $\forall y \in Y_{\text{good}}$:



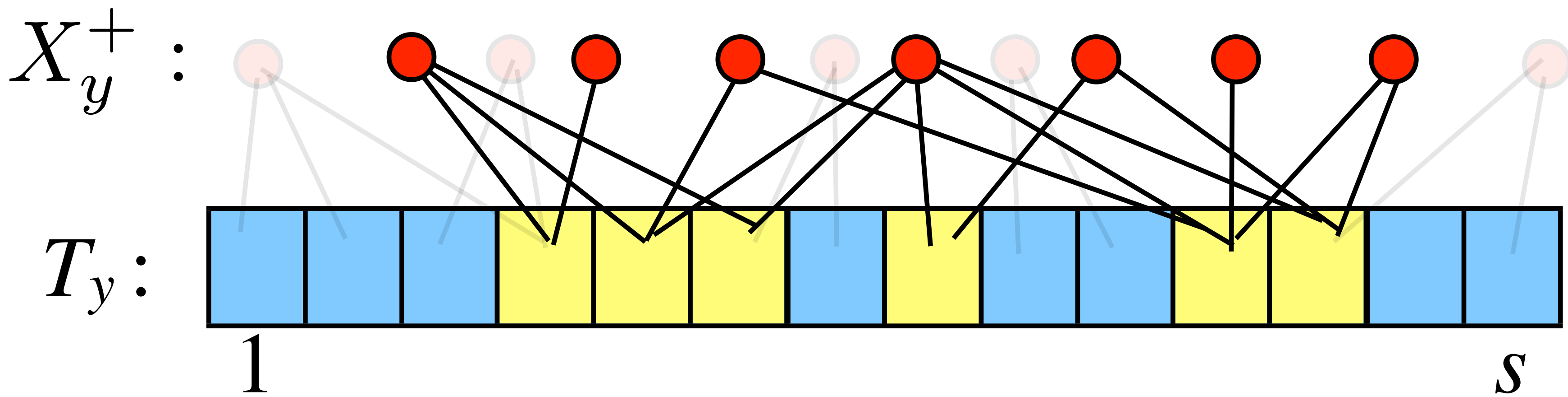
hypergraph with vertices $[s]$ and hyperedges X_y^+

average size of hyperedges $\leq 80000t$ (under μ_y^+)

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

fix $\forall y \in Y_{\text{good}}$:



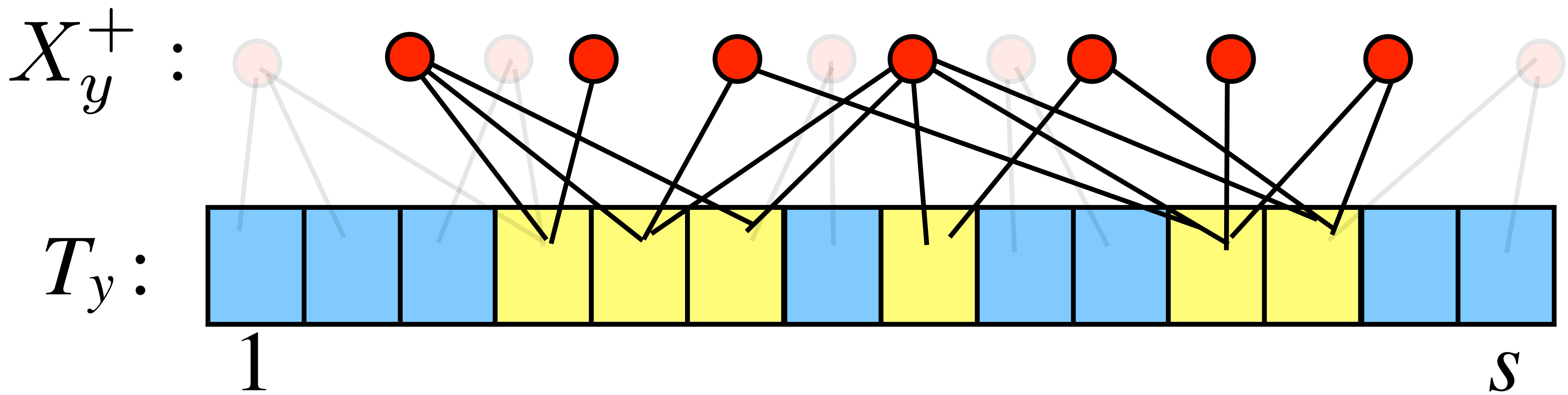
hypergraph with vertices $[s]$ and hyperedges X_y^+

average size of hyperedges $\leq 80000t$ (under μ_y^+)

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

fix $\forall y \in Y_{\text{good}}$:



hypergraph with vertices $[s]$ and hyperedges X_y^+

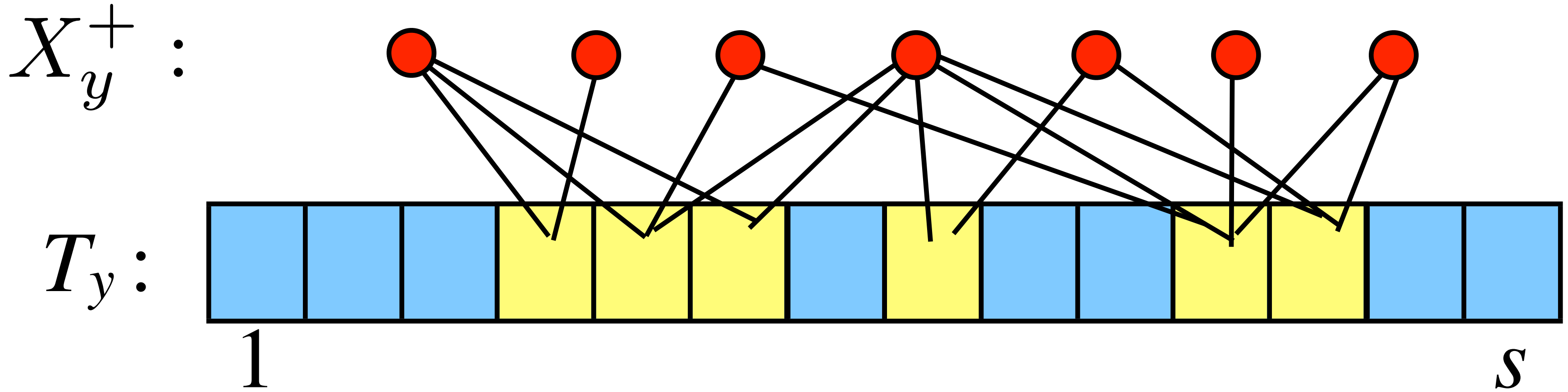
average size of hyperedges $\leq 80000t$ (under μ_y^+)

probabilistic method ⇒ $\forall\ \Delta \geq 320000t$, $\exists$ *sub*-hypergraph *induced* by Δ vertices of measure $\frac{1}{2}\left(\frac{\Delta}{2s}\right)^{80000t} \geq 2^{-O(t\log\frac{s}{\Delta})}$ (under μ_y^+)

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x}\sim\mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

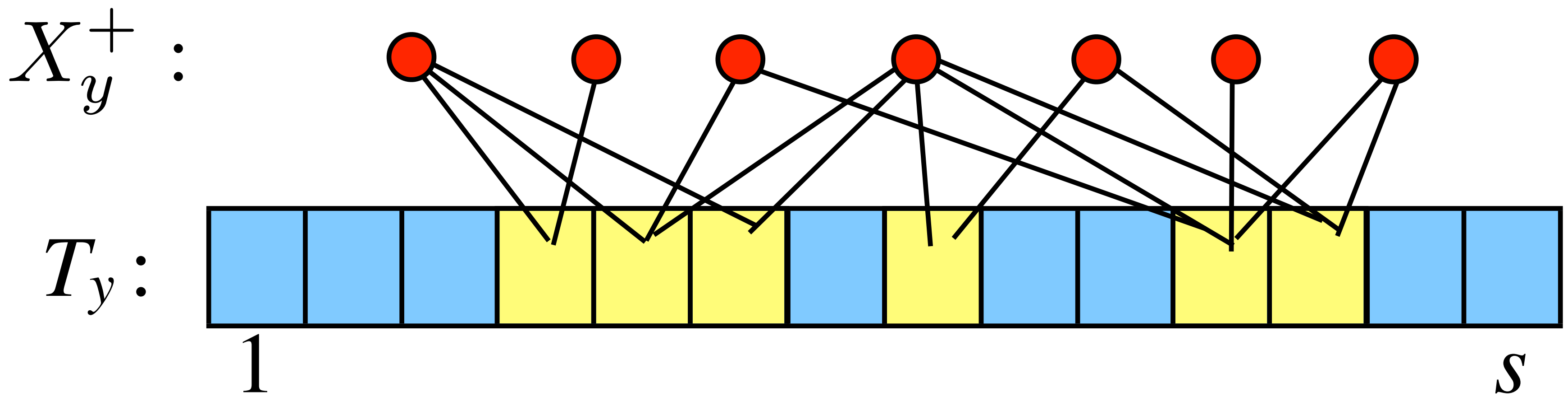
fix $\forall y \in Y_{\text{good}}$:



$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x} \sim \mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

fix $\forall y \in Y_{\text{good}}$:



$\forall\ \Delta \geq 320000t$:

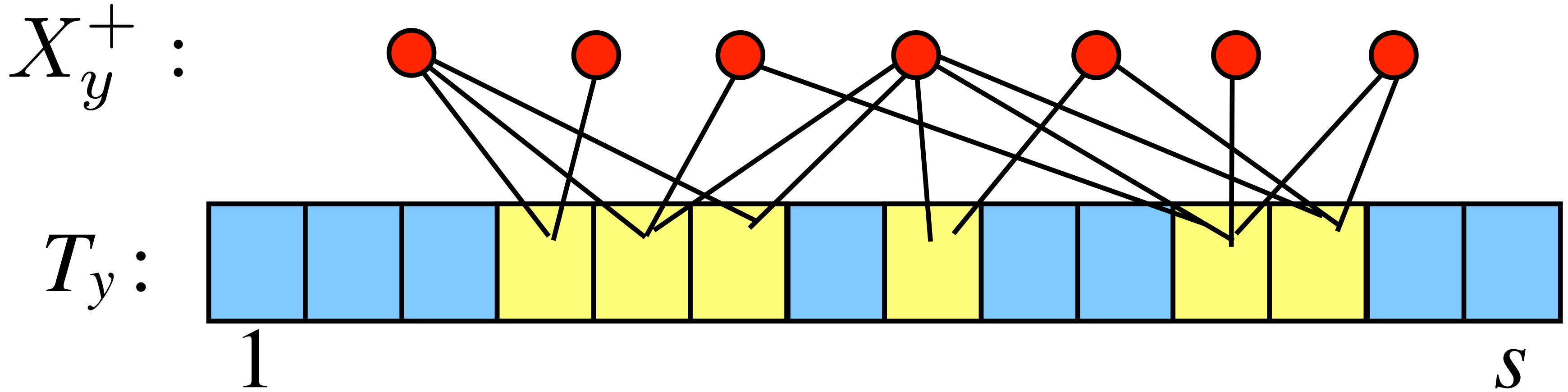
$\exists\ \Delta$ cells resolving $0.0025(\Delta/2s)^{80000t} = 2^{-O(t \log (s/\Delta))}$

fraction of positive queries (under μ)

$\exists Y_{\text{good}} \subseteq Y$ s.t. $\nu(Y_{\text{good}}) \geq 0.0025$ and $\forall y \in Y_{\text{good}}$:

- $\mu(X_y^+) \geq 0.005$; (large amount of positive queries)
- $\mathbf{E}_{\boldsymbol{x} \sim \mu_y^+}[t(\boldsymbol{x}, y)] \leq 80000t$. (bounded average cost over positive queries)

fix $\forall y \in Y_{\text{good}}$:



$\forall \ \Delta \geq 320000t$:

$\exists \ \Delta$ cells resolving $0.0025(\Delta/2s)^{80000t} = 2^{-O(t \log (s/\Delta))}$ fraction of positive queries (under μ)

good $y \longmapsto \omega$ resolving $2^{-O(t \log (s/\Delta))}$ positive queries

A New Richness Lemma

$f : X \times Y \to \{0,1\}$ distributions μ over X, ν over Y

New Richness lemma

f is 0.01-dense under $\mu\times\nu$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu$

⇒

$\forall \Delta \in [320000t, s]$,

f has 1-rectangle $A\times B$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu(B) \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

Conjunction Problem

data structure problem $f : X \times Y \to \{0, 1\}$

Conjunction Problem

data structure problem $f : X \times Y \to \{0,1\}$

$Y = Z^n$: database $y=(y_1,\ldots,y_n)$ where $y_i \in Z$

Conjunction Problem

data structure problem $f : X \times Y \to \{0,1\}$

$Y = Z^n$: database $y=(y_1,\ldots,y_n)$ where $y_i \in Z$

point-wise function: $g : X \times Z \to \{0,1\}$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Conjunction Problem

data structure problem $f : X \times Y \to \{0, 1\}$

$Y = Z^n$: database $y=(y_1,\ldots,y_n)$ where $y_i \in Z$

point-wise function: $g : X \times Z \to \{0, 1\}$

$$f(x, y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

(γ, λ)-ANN: $X = Z$ is the metric space

$$g(x, y_i) = \begin{cases} 1 & \mathrm{dist}(x, y_i) > \gamma\lambda \\ 0 & \mathrm{dist}(x, y_i) \leq \lambda \\ * & \text{otherwise} \end{cases}$$

Conjunction Problem

data structure problem $f : X \times Y \to \{0,1\}$

$Y = Z^n$: database $y=(y_1,\ldots,y_n)$ where $y_i \in Z$

point-wise function: $g : X \times Z \to \{0,1\}$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

(γ, λ)-ANN: $X = Z$ is the metric space

$$g(x,y_i) = \begin{cases} 1 & \text{dist}(x,y_i) > \gamma\lambda \\ 0 & \text{dist}(x,y_i) \le \lambda \\ * & \text{otherwise} \end{cases}$$

Examples: partial match, membership, range query, ...

data structure problem $f : X \times Z^n \to \{0, 1\}$

$$f(x, y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

with point-wise function $g : X \times Z \to \{0, 1\}$

data structure problem $f : X \times Z^n \rightarrow \{0,1\}$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

with point-wise function $g : X \times Z \rightarrow \{0,1\}$

distributions μ over X, ν over Z, ν^n over $Y = Z^n$

data structure problem $f : X \times Z^n \to \{0,1\}$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

with point-wise function $g : X \times Z \to \{0,1\}$

distributions μ over X, ν over Z, ν^n over $Y = Z^n$

Theorem

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$;
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

If f has an average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

$$\Rightarrow \quad t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right) \quad \text{or} \quad t = \Omega\left(\frac{n \log \Psi}{w + \log s}\right)$$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu \times \nu$;
- g has no 1-rectangle $A \times B$ with $\mu(A) \geq 1/\Phi$ and $\nu(B) \geq 1/\Psi$.

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$; ⇒ (union bound) **f is 0.01-dense under $\mu\times\nu^n$**
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$; ⇒ (union bound) f is 0.01-dense under $\mu\times\nu^n$
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

⇒ $\forall\Delta\in[320000t,s]$,

f has 1-rectangle $A\times B'$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t\log(s/\Delta))} \\ \nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)} \end{cases}$$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$;
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

union bound ⇒ f is 0.01-dense under $\mu\times\nu^n$

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

⇒ $\forall\Delta\in[320000t, s]$, f has 1-rectangle $A\times B'$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t\log(s/\Delta))} \\ \nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)} \end{cases}$$

choose $\Delta = O\left(\frac{n\log\Psi}{w}\right)\cap\Omega\left(\frac{n\log\Psi}{w+\log s}\right)$ to satisfy

$$\nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)} > 1/\Psi^n$$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$; ⇒ (union bound) f is 0.01-dense under $\mu\times\nu^n$
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

⇒ $\forall \Delta\in[320000t, s]$, f has 1-rectangle $A\times B'$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t\log(s/\Delta))} \\ \nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta) + \Delta w)} \end{cases}$$

choose $\Delta = O\left(\frac{n\log\Psi}{w}\right) \cap \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$ to satisfy

$$\nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta) + \Delta w)} > 1/\Psi^n$$

case 1: $\Delta < 320000t$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$; ⇨ (union bound) **f is 0.01-dense under $\mu\times\nu^n$**
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

⇨ $\forall \Delta \in [320000t, s]$, f has 1-rectangle $A\times B'$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t\log(s/\Delta))} \\ \nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta) + \Delta w)} \end{cases}$$

choose $\Delta = O\left(\frac{n\log\Psi}{w}\right) \cap \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$ to satisfy

$$\nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta) + \Delta w)} > 1/\Psi^n$$

case 1: $\Delta < 320000t$ ⇨ $t = \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$;
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

union bound $\Rightarrow$ f is 0.01-dense under $\mu\times\nu^n$

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

$\Rightarrow$ $\forall\Delta\in[320000t, s]$, f has 1-rectangle $A\times B'$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t \log (s/\Delta))} \\ \nu^n(B') \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} \end{cases}$$

choose $\Delta = O\left(\frac{n\log\Psi}{w}\right) \cap \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$ to satisfy

$$\nu^n(B') \geq 2^{-O(\Delta \log (s/\Delta) + \Delta w)} > 1/\Psi^n$$

case 1: $\Delta < 320000t$ $\Rightarrow$ $t = \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$

case 2: otherwise

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$; → (union bound) f is 0.01-dense under $\mu\times\nu^n$
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

⟹ $\forall\Delta\in[320000t, s]$, f has 1-rectangle $A\times B'$ with

$$\begin{cases} \mu(A) \geq 2^{-O(t\log(s/\Delta))} \\ \nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)} \end{cases}$$

choose $\Delta = O\left(\frac{n\log\Psi}{w}\right)\cap\Omega\left(\frac{n\log\Psi}{w+\log s}\right)$ to satisfy

$$\nu^n(B') \geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)} > 1/\Psi^n$$

case 1: $\Delta < 320000t$ ⟹ $t = \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$

case 2: otherwise ⟹ $1/\Phi > \mu(A) \geq 2^{-O(t\log(s/\Delta))}$

$$f(x,y)=\bigwedge_{i=1}^{n} g(x,y_i)$$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$;
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

union bound ⇒ f is 0.01-dense under $\mu\times\nu^n$

New Richness lemma

f is 0.01-dense under $\mu\times\nu^n$

f has average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

⇒ $\forall\Delta\in[320000t, s]$, f has 1-rectangle $A\times B'$ with

$$\begin{cases}\mu(A)\geq 2^{-O(t\log(s/\Delta))}\\ \nu^n(B')\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}\end{cases}$$

choose $\Delta=O\left(\frac{n\log\Psi}{w}\right)\cap\Omega\left(\frac{n\log\Psi}{w+\log s}\right)$ to satisfy

$$\nu^n(B')\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)}>1/\Psi^n$$

case 1: $\Delta<320000t$ ⇒ $t=\Omega\left(\frac{n\log\Psi}{w+\log s}\right)$

case 2: otherwise ⇒ $1/\Phi>\mu(A)\geq 2^{-O(t\log(s/\Delta))}$ ⇒ $t=\Omega\left(\frac{\log\Phi}{\log\frac{sw}{n\log\Psi}}\right)$

data structure problem $f : X \times Z^n \to \{0,1\}$

$$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

with point-wise function $g : X \times Z \to \{0,1\}$

distributions μ over X, ν over Z, ν^n over $Y = Z^n$

Theorem

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu\times\nu$;
- g has no 1-rectangle $A\times B$ with $\mu(A)\geq 1/\Phi$ and $\nu(B)\geq 1/\Psi$.

If f has an average-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

$$\Rightarrow \quad t = \Omega\left(\frac{\log\Phi}{\log\frac{sw}{n\log\Psi}}\right) \quad \text{or} \quad t = \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$$

(γ, λ)-ANN: $$f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$$

$$g(x, y_i) = \begin{cases} 1 & \text{dist}(x, y_i) > \gamma\lambda \\ 0 & \text{dist}(x, y_i) \le \lambda \\ * & \text{otherwise} \end{cases}$$

distributions μ over X, μ over X, μ^n over $Y = X^n$

Theorem

Assume:

- density of 0s in g is $\le 0.99/n$ under $\mu \times \mu$;
- g has no 1-rectangle $A \times B$ with $\mu(A) \ge 1/\Phi$ and $\mu(B) \ge 1/\Psi$.

If f has an average-case (s,w,t)-cell-probing scheme under $\mu \times \mu^n$

$$t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right) \quad \text{or} \quad t = \Omega\left(\frac{n \log \Psi}{w + \log s}\right)$$

(γ, λ)-ANN: $f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$

$$g(x, y_i) = \begin{cases} 1 & \text{dist}(x, y_i) > \gamma\lambda \\ 0 & \text{dist}(x, y_i) \le \lambda \\ * & \text{otherwise} \end{cases}$$

distributions μ over X, μ over X, μ^n over $Y = X^n$

Theorem

Assume:

- density of 0s in g is $\le 0.99/n$ under $\mu\times\mu$; (weakly independence)
- g has no 1-rectangle $A\times B$ with $\mu(A)\ge 1/\Phi$ and $\mu(B)\ge 1/\Psi$.

If f has an average-case (s,w,t)-cell-probing scheme under $\mu\times\mu^n$

$\Rightarrow$ $t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right)$ or $t = \Omega\left(\frac{n \log \Psi}{w + \log s}\right)$

(γ, λ)-ANN: $f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$

$$g(x, y_i) = \begin{cases} 1 & \text{dist}(x, y_i) > \gamma\lambda \\ 0 & \text{dist}(x, y_i) \le \lambda \\ * & \text{otherwise} \end{cases}$$

distributions μ over X, μ over X, μ^n over $Y = X^n$

Theorem

Assume:

- density of 0s in g is $\le 0.99/n$ under $\mu\times\mu$; (weakly independence)
- g has no 1-rectangle $A\times B$ with $\mu(A)\ge 1/\Phi$ and $\mu(B)\ge 1/\Psi$. $((\Phi, \Psi)$-expanding$)$

If f has an average-case (s,w,t)-cell-probing scheme under $\mu\times\mu^n$

$\Rightarrow$ $t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right)$ or $t = \Omega\left(\frac{n \log \Psi}{w + \log s}\right)$

(γ, λ)-ANN: $\quad f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$

$$g(x, y_i) = \begin{cases} 1 & \mathrm{dist}(x, y_i) > \gamma\lambda \\ 0 & \mathrm{dist}(x, y_i) \leq \lambda \\ * & \text{otherwise} \end{cases}$$

distributions μ over X, μ over X, μ^n over $Y = X^n$

Assume:

- density of 0s in g is $\leq 0.99/n$ under $\mu \times \mu$;
- g has no 1-rectangle $A \times B$ with $\mu(A) \geq 1/\Phi$ and $\mu(B) \geq 1/\Psi$.

(γ, λ)-ANN: $f(x,y) = \bigwedge_{i=1}^{n} g(x, y_i)$

$$g(x, y_i) = \begin{cases} 1 & \text{dist}(x, y_i) > \gamma\lambda \\ 0 & \text{dist}(x, y_i) \le \lambda \\ * & \text{otherwise} \end{cases}$$

distributions μ over X, μ over X, μ^n over $Y = X^n$

Assume:

- density of 0s in g is $\le 0.99/n$ under $\mu \times \mu$;
- g has no 1-rectangle $A \times B$ with $\mu(A) \ge 1/\Phi$ and $\mu(B) \ge 1/\Psi$.

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

$$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$

(γ,λ)-ANN: $f(x,y)=\bigwedge_{i=1}^{n} g(x,y_i)$

$$g(x,y_i)=\begin{cases}1 & \mathrm{dist}(x,y_i)>\gamma\lambda\\ 0 & \mathrm{dist}(x,y_i)\le\lambda\\ * & \text{otherwise}\end{cases}$$

distributions μ over X, μ over X, μ^n over $Y=X^n$

Assume:

- density of 0s in g is $\le 0.99/n$ under $\mu\times\mu$;
- g has no 1-rectangle $A\times B$ with $\mu(A)\ge 1/\Phi$ and $\mu(B)\ge 1/\Psi$.

- $\gamma\lambda$-neighborhoods are weakly independent under μ:
  $\mu(N_{\gamma\lambda}(x)) < 0.99/n$ for $\forall x\in X$
- λ-neighborhoods are (Φ,Ψ)-expanding under μ:
  $\forall A\subseteq X$, $\mu(A)\ge 1/\Phi \Rightarrow \mu(N_\lambda(A))\ge 1\text{-}1/\Psi$

metric space (X,dist), distribution μ over X:

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

$$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$

- λ-neighborhoods are (Φ,Ψ)-expanding under μ:

$$\forall A \subseteq X,\ \mu(A) \geq 1/\Phi \Rightarrow \mu(N_\lambda(A)) \geq 1-1/\Psi$$

Theorem

Assume:

- $\gamma\lambda$-neighborhoods are weakly independent under μ;
- λ-neighborhoods are (Φ,Ψ)-expanding under μ.

If (γ,λ)-*ANN* has avg.-case (s,w,t)-cell-probing scheme under $\mu\times\nu^n$

$$t = \Omega\left(\frac{\log\Phi}{\log\frac{sw}{n\log\Psi}}\right) \quad \text{or} \quad t = \Omega\left(\frac{n\log\Psi}{w+\log s}\right)$$

Hamming space $X=\{0,1\}^d$, uniform distribution μ over X:

Hamming space $X=\{0,1\}^d$, uniform distribution μ over X:

choose $\gamma\lambda = \frac{d}{2} - \sqrt{2d\ln(2n)}$

Hamming space $X=\{0,1\}^d$, uniform distribution μ over X:

choose $\gamma\lambda = \frac{d}{2} - \sqrt{2d\ln(2n)}$

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

  $\mu(N_{\gamma\lambda}(x)) < 0.99/n$ for $\forall x \in X$

Hamming space $X=\{0,1\}^d$, uniform distribution μ over X:

choose $\gamma\lambda = \frac{d}{2} - \sqrt{2d\ln(2n)}$

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

$$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$

Harper's Isoperimetric inequality:

$$\forall A \subseteq X,\ \mu(A) \geq \mu(N_r(\mathbf{\underline{0}})) \Rightarrow \mu(N_\lambda(A)) \geq \mu(N_{r+\lambda}(\mathbf{\underline{0}}))$$

"Hamming balls have the smallest vertex-expansion."

Hamming space $X=\{0,1\}^d$, uniform distribution μ over X:

choose $\gamma\lambda = \frac{d}{2} - \sqrt{2d\ln(2n)}$

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

$$\mu(N_{\gamma\lambda}(x)) < 0.99/n \text{ for } \forall x \in X$$

Harper's Isoperimetric inequality:

$$\forall A \subseteq X,\ \mu(A) \geq \mu(N_r(\underline{\mathbf{0}})) \Rightarrow \mu(N_\lambda(A)) \geq \mu(N_{r+\lambda}(\underline{\mathbf{0}}))$$

"Hamming balls have the smallest vertex-expansion."

- λ-neighborhoods are $(2^{\Theta(d)}, 2^{\Theta(d)})$-expanding under μ:

$$\forall A \subseteq X,\ \mu(A) \geq 2^{-\Theta(d)} \Rightarrow \mu(N_\lambda(A)) \geq 1-2^{-\Theta(d)}$$

Hamming space $X=\{0,1\}^d$, uniform distribution μ over X:

choose $$\gamma\lambda = \frac{d}{2} - \sqrt{2d\ln(2n)}$$

- $\gamma\lambda$-neighborhoods are weakly independent under μ:

  $\mu(N_{\gamma\lambda}(x)) < 0.99/n$ for $\forall x \in X$

Harper's Isoperimetric inequality:

$$\forall A \subseteq X,\ \mu(A) \geq \mu(N_r(\mathbf{\underline{0}})) \Rightarrow \mu(N_\lambda(A)) \geq \mu(N_{r+\lambda}(\mathbf{\underline{0}}))$$

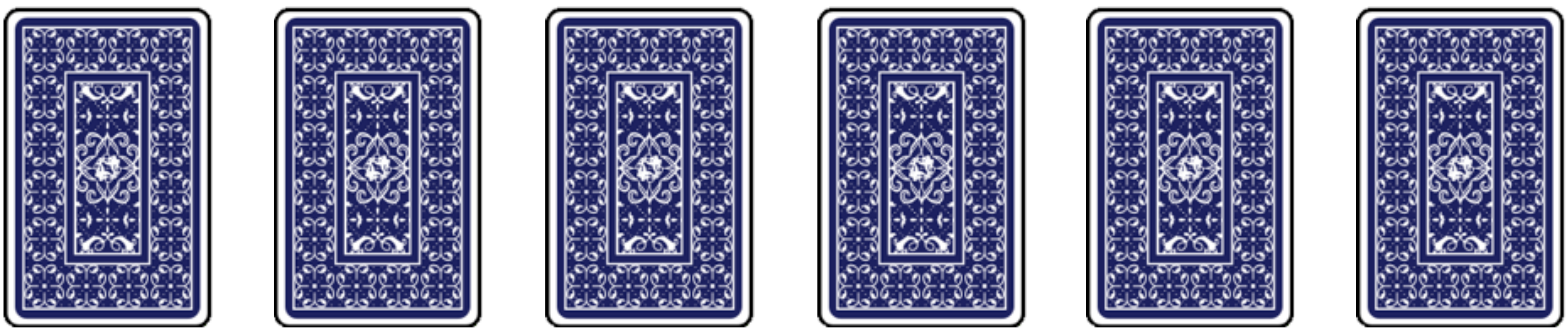
"Hamming balls have the smallest vertex-expansion."

- λ-neighborhoods are $(2^{\Theta(d)}, 2^{\Theta(d)})$-expanding under μ:

  $\forall A \subseteq X,\ \mu(A) \geq 2^{-\Theta(d)} \Rightarrow \mu(N_\lambda(A)) \geq 1-2^{-\Theta(d)}$

$$t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$$ or $$t = \Omega\left(\frac{nd}{w + \log s}\right)$$

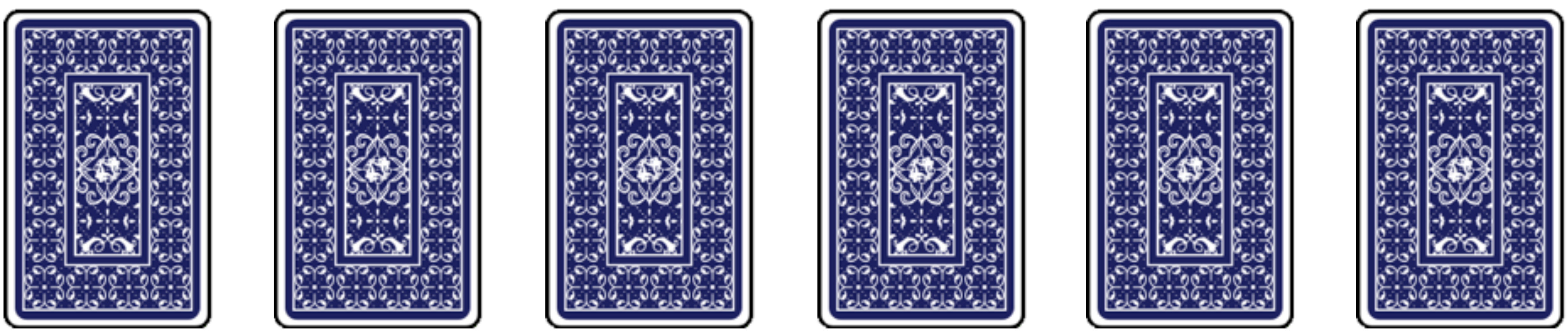
Certificates in Data Structures



certificate: the cells whose contents *uniquely identify* the answer

Certificates in Data Structures

for one-dimensional nearest neighbor search
(predecessor search)

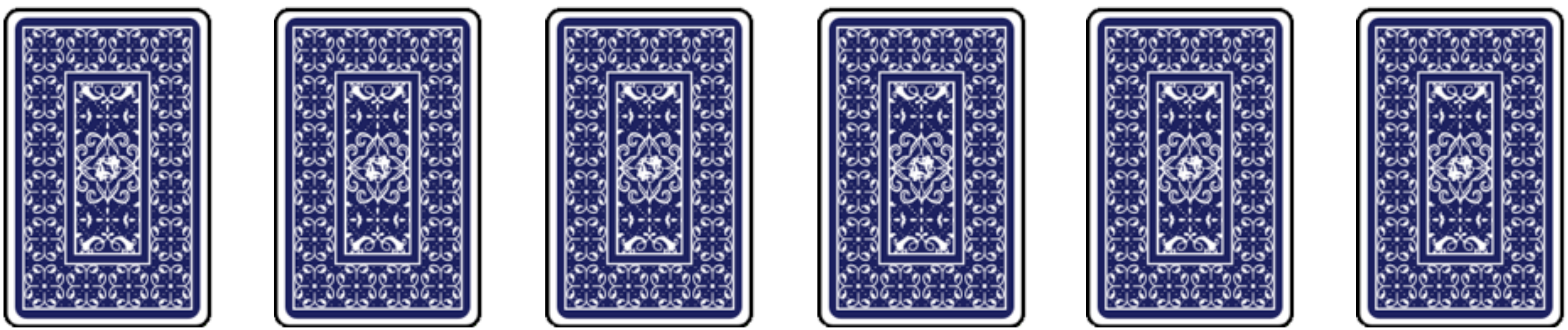


certificate: the cells whose contents *uniquely identify* the answer

Certificates in Data Structures

for one-dimensional nearest neighbor search
(predecessor search)

assuming the data structure is the sorted table:

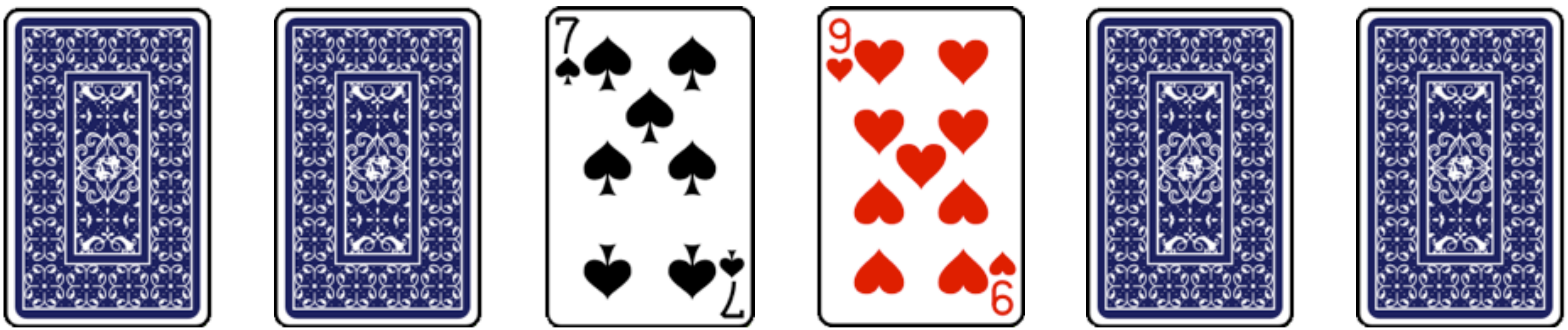


certificate: the cells whose contents *uniquely identify* the answer

Certificates in Data Structures

for one-dimensional nearest neighbor search
(predecessor search)

assuming the data structure is the sorted table:

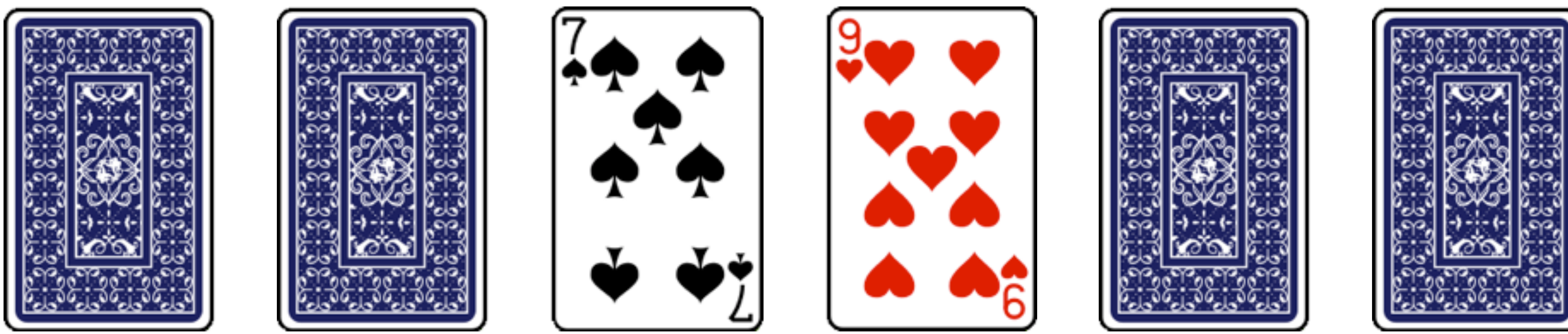


certificate: the cells whose contents *uniquely identify* the answer

Certificates in Data Structures

for one-dimensional nearest neighbor search
(predecessor search)

assuming the data structure is the sorted table:



certifies the nearest neighbor of "8.5"

certificate: the cells whose contents *uniquely identify* the answer

Lower Bounds for Hamming NNS

Hamming space $X = \{0,1\}^d$ database $\boldsymbol{y} \in X^n$

time: t cell-probes; space: s cells, each of w bits

	deterministic	randomized
exact	average-case: $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$	average-case: $t = \Omega\left(\frac{d}{\log s}\right)$ [Barkol Rabani 2000] worst-case: $t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
approx.	average-case: $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$	worst-case, search problem: $t = \Theta\left(\frac{\log\log d}{\log\log\log d}\right)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004] average-case: $t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]

Thank you!