Simple Average-case Lower Bounds *for* Approximate Near-Neighbor *from* Isoperimetric Inequalities

Yitong Yin
Nanjing University

Nearest Neighbor Search
(*NNS*)

metric space (X,dist)

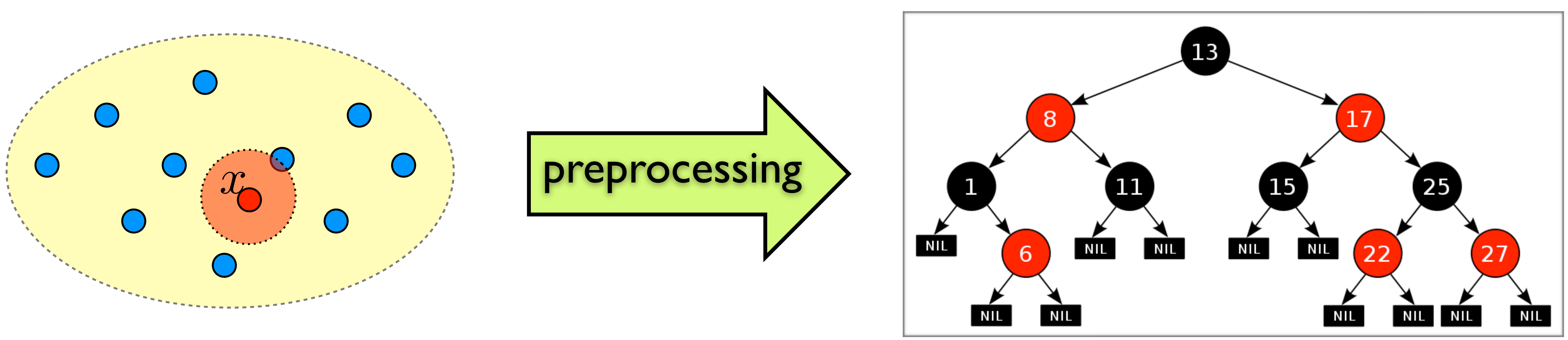
query $x \in X$

↓ access

data structure

database

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



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 $x \in X$

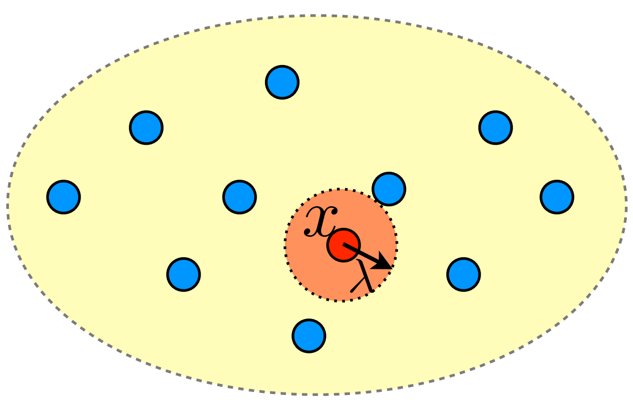
↓ access

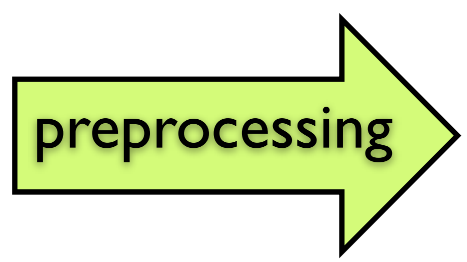
database

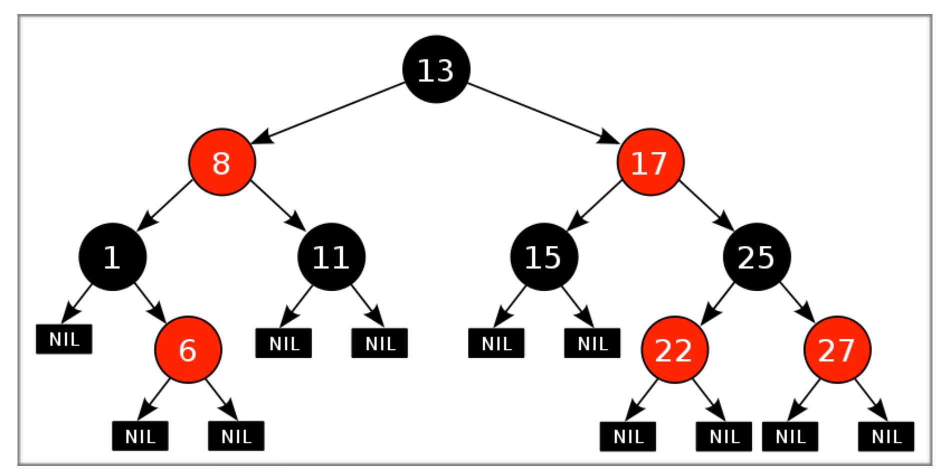
$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure

radius λ







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

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

Approximate Near Neighbor
(ANN)

metric space (X,dist)

query $x \in X$

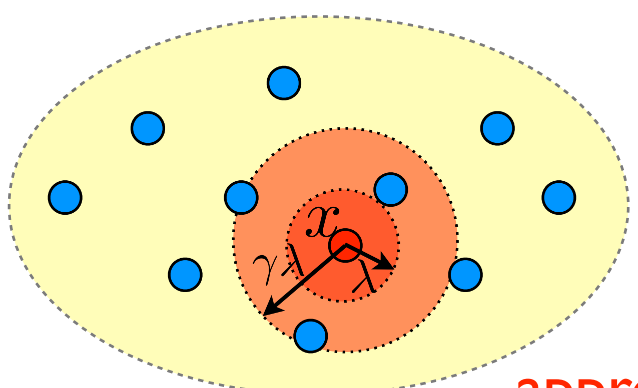
access

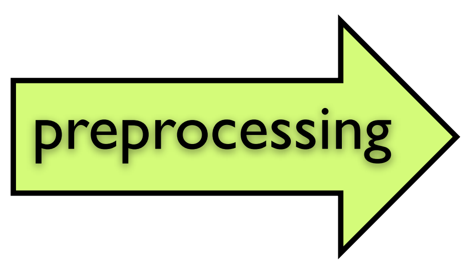
database

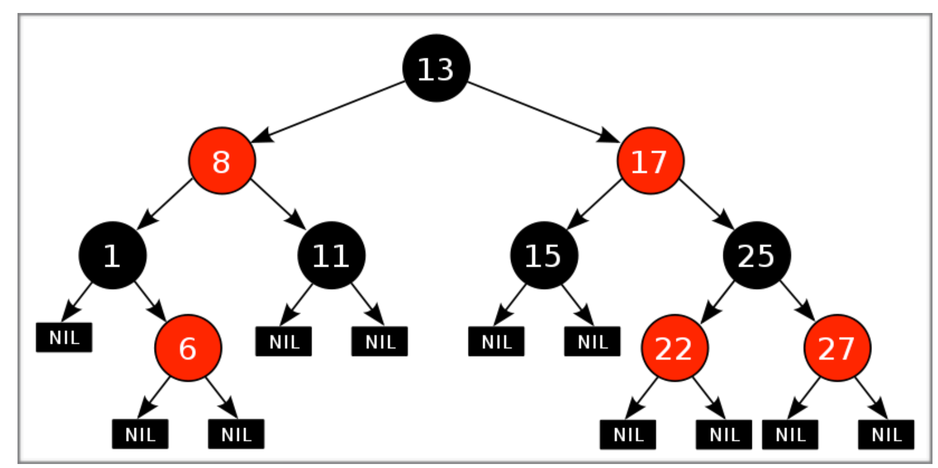
$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure

radius λ







approximation ratio $\gamma \geq 1$

(γ, λ)-ANN: answer "yes" if $\exists y_i$ that is $\leq\lambda$-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)

query $x \in X$

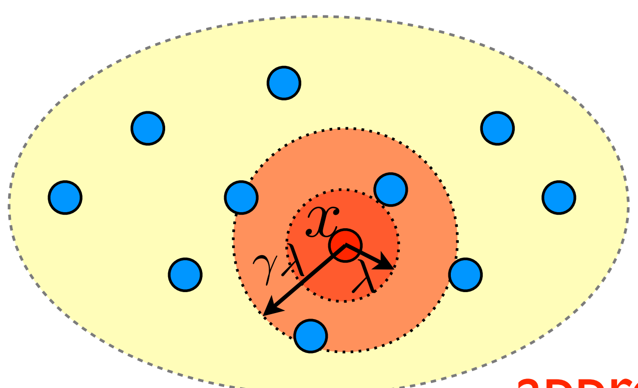
access

database

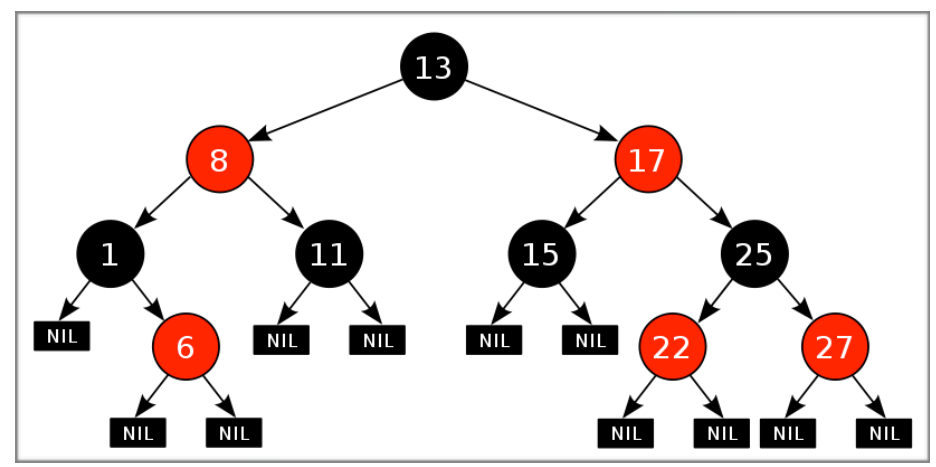
$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in X^n$

data structure

radius λ



preprocessing



approximation ratio $\gamma \geq 1$

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

Hamming distance

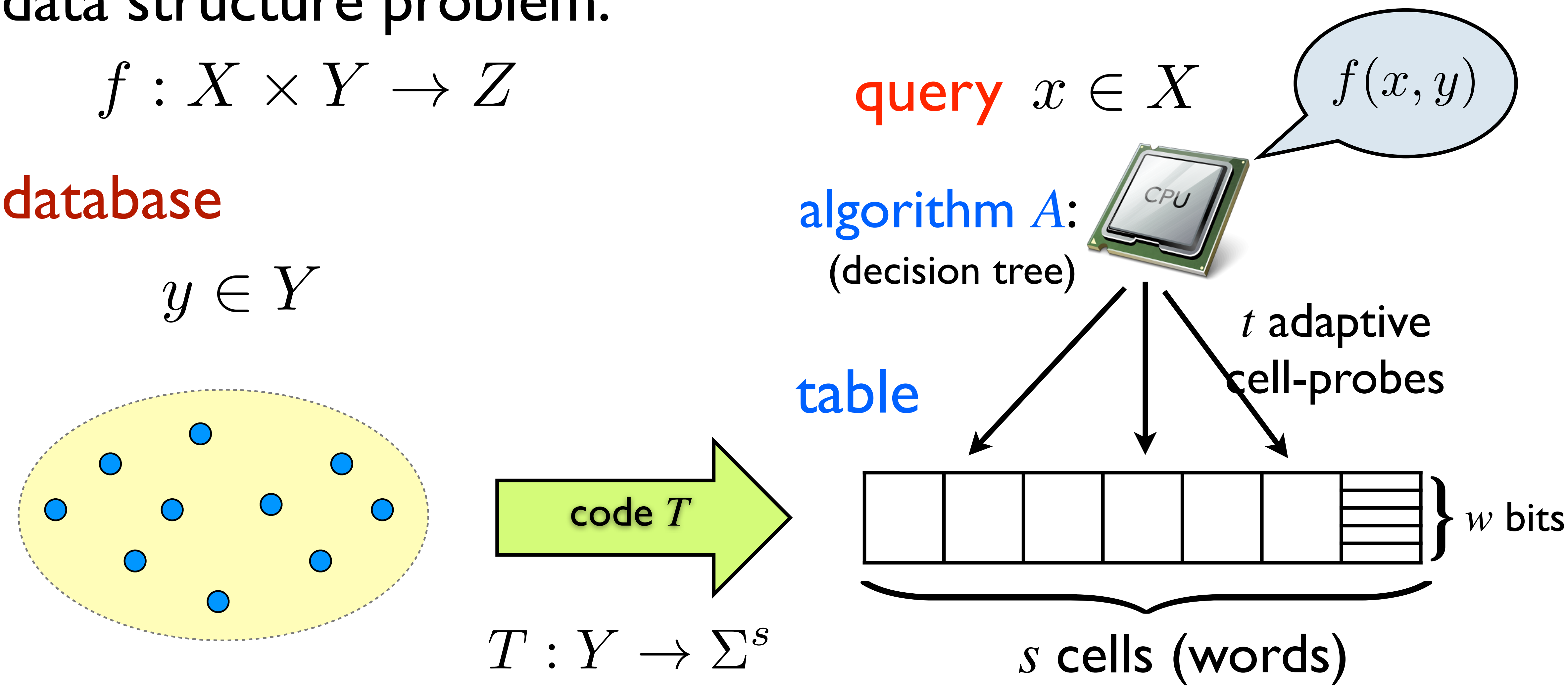
$$100 \log n < d < n^{o(1)}$$

Curse of dimensionality!

Cell-Probe Model

data structure problem:

$f : X \times Y \to Z$



$T : Y \to \Sigma^s$

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

protocol: the pair (A, T)

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

Near-Neighbor Lower Bounds

Hamming space $X = \{0,1\}^d$ database size: n / $d = \Theta(\log n)$

time: t cell-probes; linear space / space: s cells, each of w bits; $s = O(n)$, $w = O(d)$

Approximate Near-Neighbor (ANN)		Randomized *Exact* Near-Neighbor (*RENN*)
Deterministic	Randomized	
$t = \Omega\left(\frac{d}{\log s}\right)$ / $t = \Omega(1)$ [Miltersen *et al.* 1995] [Liu 2004]	$t = O(1)$ for $s = \mathrm{poly}(n)$ [Chakrabarti Regev 2004]	$t = \Omega\left(\frac{d}{\log s}\right)$ / $t = \Omega(1)$ [Borodin Ostrovsky Rabani 1999] [Barkol Rabani 2000]
$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ / $t = \Omega\left(\frac{\log n}{\log\log n}\right)$ [Pătraşcu Thorup 2006]	$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ / $t = \Omega\left(\frac{\log n}{\log\log n}\right)$ [Panigrahy Talwar Wieder 2008, 2010]	$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ / $t = \Omega\left(\frac{\log n}{\log\log n}\right)$ [Pătraşcu Thorup 2006]
$t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ / $t = \Omega(\log n)$ [Wang Y. 2014]		

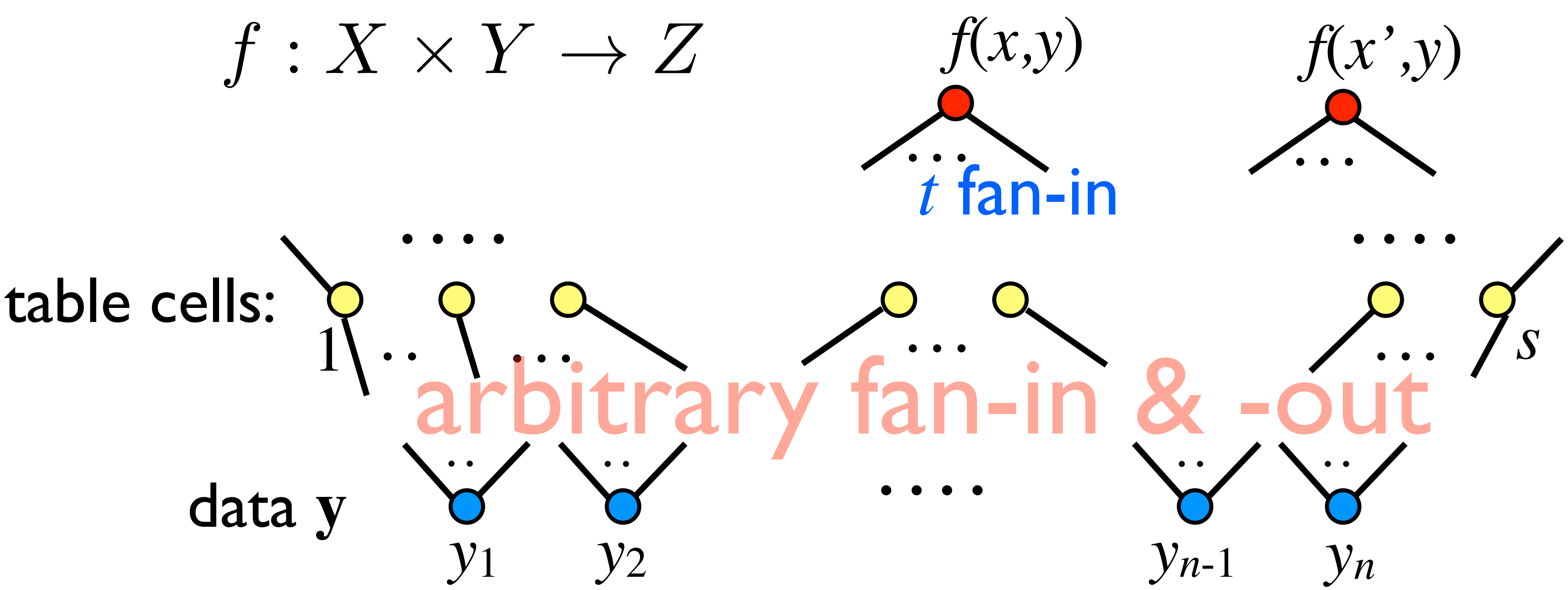
- matches the highest known lower bounds for any data structure problems: Polynomial Evaluation [Larsen'12], ball-inheritance (range reporting) [Grønlund, Larsen'16]

Why are data structure lower bounds so difficult?

- (Observed by [Miltersen *et al.* 1995]) An $\omega(\log n)$ cell-probe lower bound on polynomial space for *any* function in **P** would prove **P** $\not\subseteq$ linear-time poly-size Boolean branching programs. (Solved in [Ajtai 1999])

- (Observed by [Brody, Larsen 2012]) Even *non-adaptive* data structures are circuits with arbitrary gates of depth 2:

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



Near-Neighbor Lower Bounds

Hamming space $X = \{0,1\}^d$ database size: n

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

Approximate Near-Neighbor (ANN)		Randomized *Exact* Near-Neighbor (*RENN*)
Deterministic	**Randomized**	
$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] [Liu 2004]	$t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Borodin Ostrovsky Rabani 1999] [Barkol Rabani 2000]
$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]		$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]
$t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]		

Average-Case Lower Bounds

- Hard distribution: [Barkol Rabani 2000] [Liu 2004] [PTW'08 '10]
  - database: $y_1,\ldots,y_n \in \{0,1\}^d$ *i.i.d. uniform*
  - query: uniform and independent $x \in \{0,1\}^d$
- *Expected* cell-probe complexity:
  - $\mathbf{E}_{(x,y)}$[# of cell-probes to resolve query x on database y]
- "Curse of dimensionality" should hold on average.
- In *data-dependent* LSH [Andoni Razenshteyn 2015]: a key step is to solve the problem on random input.

Average-Case Lower Bounds

Hamming space $X = \{0,1\}^d$ database size: n

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

Approximate Near-Neighbor (ANN)		Randomized *Exact* Near-Neighbor (*RENN*)
Deterministic	Randomized	
$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] [Liu 2004]	$t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Borodin Ostrovsky Rabani 1999] [Barkol Rabani 2000]
~~$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]~~		~~$t = \Omega\left(\frac{d}{\log \frac{sw}{n}}\right)$ [Pătraşcu Thorup 2006]~~
~~$t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$ [Wang Y. 2014]~~		

Average-Case Lower Bounds

Hamming space $X = \{0,1\}^d$ database size: n

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

Approximate Near-Neighbor (ANN)		Randomized *Exact* Near-Neighbor (*RENN*)
Deterministic	Randomized	
$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] [Liu 2004]	$t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Borodin Ostrovsky Rabani 1999] [Barkol Rabani 2000]
our result: $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$		

Metric Expansion

[Panigrahy Talwar Wieder 2010]

metric space (X,dist)

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

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

probability distribution μ over X

- 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:

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

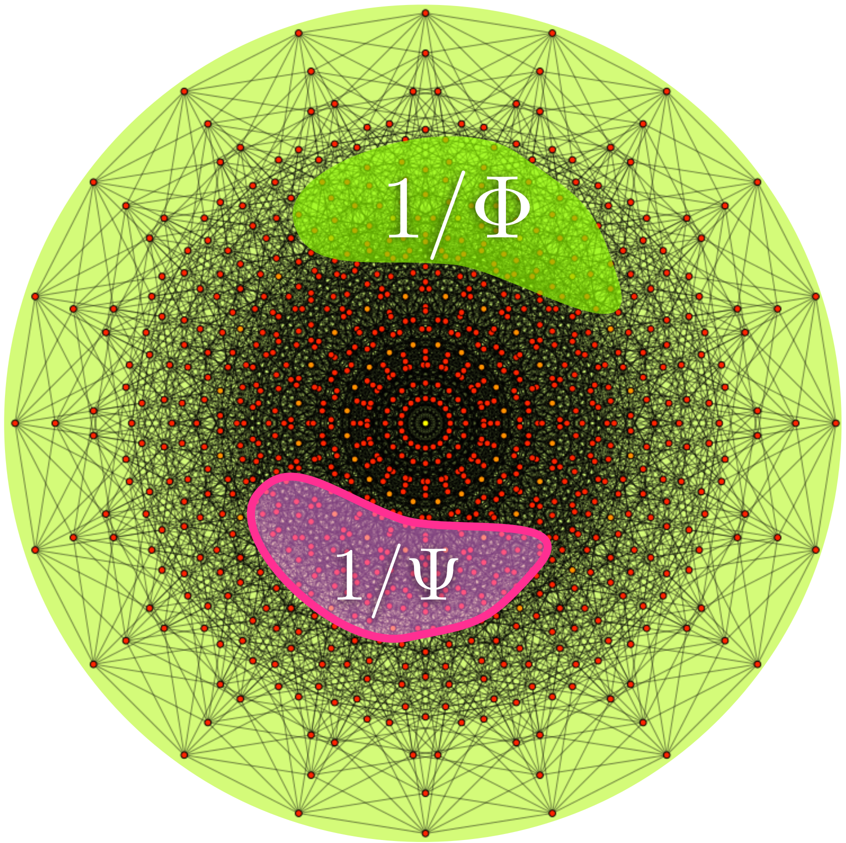
Metric Expansion

[Panigrahy Talwar Wieder 2010]

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

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

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



vertex expansion, "blow-up" effect

Main Theorem:

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

- $\gamma\lambda$-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:
  $$\forall A \subseteq X \text{ that } \mu(A) \geq 1/\Phi \Rightarrow \mu(N_\lambda(A)) \geq 1\text{-}1/\Psi$$

$\forall$ *deterministic* algorithm that makes t cell-probes *in expectation* on a table of size s cells, each of w bits (assuming w+log $s < n$ / log Φ), under the *input distribution*:

database $y=(y_1, y_2,\ldots,y_n)$ where $y_1, y_2,\ldots,y_n \sim \mu$, *i.i.d.*

query $x \sim \mu$, *independently*

$$t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right)$$

Main Theorem:

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

- $\gamma\lambda$-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:

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

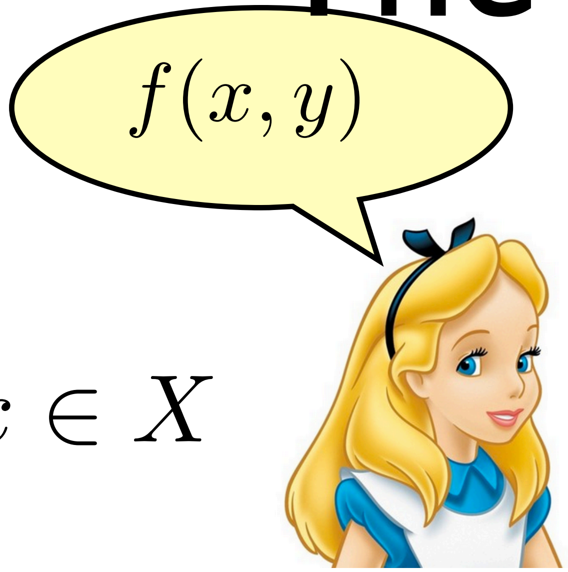
$\forall$ *deterministic* algorithm that makes t cell-probes *in expectation* on a table of size s cells, each of w bits (assuming w+log $s < n$ / log Φ), under the *input distribution*:

database $y=(y_1, y_2,\ldots,y_n)$ where $y_1, y_2,\ldots,y_n \sim \mu$, *i.i.d.*

query $x \sim \mu$, *independently*

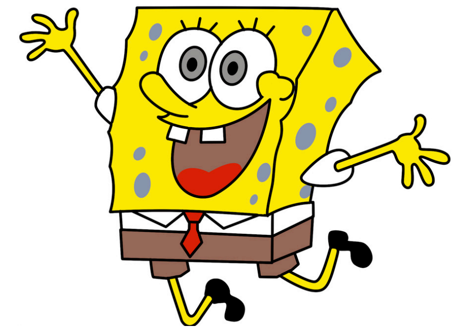
$$t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right)$$

The *Richness* Lemma



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

$x \in X$ — $t \log s$ → / ← tw — $y \in Y$



cell-probing algorithm

table (s cells, each of w bits)

distributions μ over X, ν over Y

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

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

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

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

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

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

$\Rightarrow$

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

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

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

metric space (X,dist), query $x\in X$, database $y=(y_1,\ldots,y_n)\in X^n$

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

where

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

Other examples: partial match, membership, range query, ...

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

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

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

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

⇒ density of 0s in g is $\leq 0.99/n$ under $\mu\times\mu$ ⇒ f is 0.01-dense under $\mu\times\mu^n$

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

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

⇒ g does not have 1-rectangle $A\times C$ with $\mu(A)>1/\Phi$ and $\mu(C)>1/\Psi$

⇒ f does not have 1-rectangle $A\times B$ with $\mu(A)>1/\Phi$ and $\mu^n(B)>1/\Psi^n$

choose $\Delta = O\left(\frac{n\log\Psi}{w}\right)$ so that $\mu^n(B)\geq 2^{-O(\Delta\log(s/\Delta)+\Delta w)} > 1/\Psi^n$

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

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

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

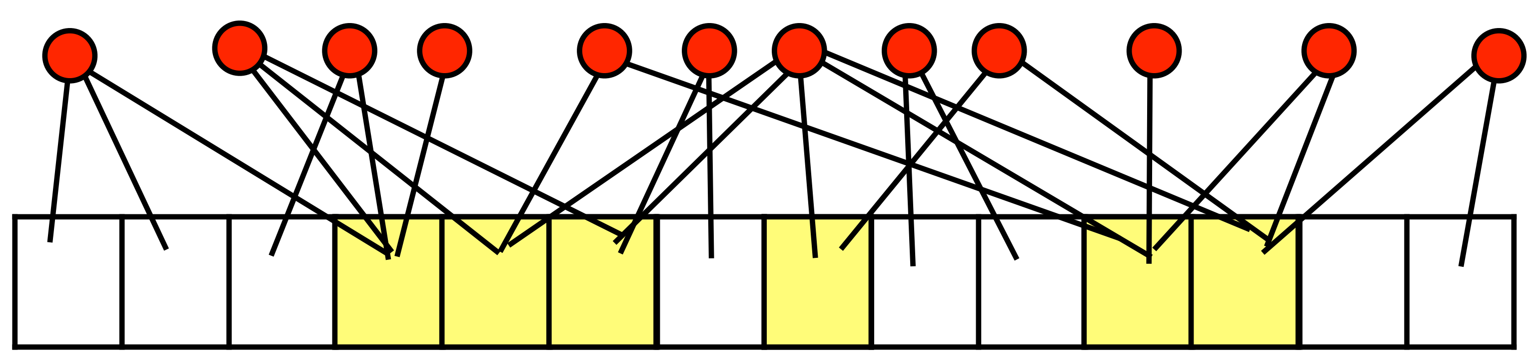
$\Downarrow$

≥ 0.0025-fraction (under ν) of databases $y\in Y$ are "good":

s.t. $\forall$ good database y, $\begin{cases}\geq 0.005\text{-fraction of queries } x\in X \text{ are positive}\\ \text{avg. cell-probes for positive queries} \leq 80000t\end{cases}$

positive queries:

T_y :



$\Rightarrow$ $\exists\ \Delta$ 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 $\mu\times\nu$

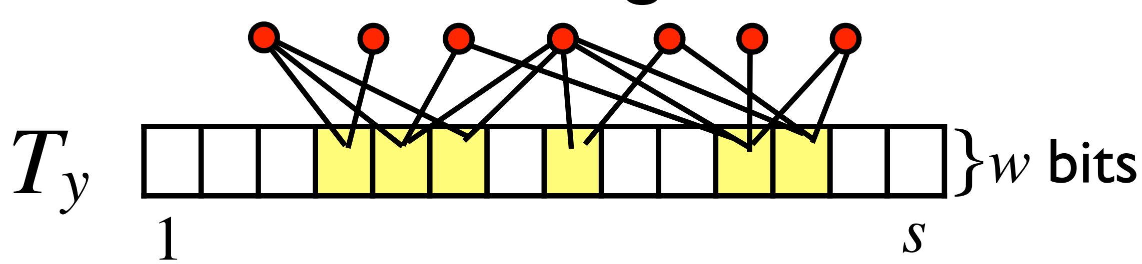
$\Rightarrow$ $\forall\,\Delta\in[320000t, s]$, f has 1-rectangle $A\times B$ with

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

≥ 0.0025-fraction (under ν) of databases $y\in Y$ are "good":

s.t. $\forall$ good database 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,

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

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

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

Main Theorem:

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

- $\gamma\lambda$-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:

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

$\forall$ *deterministic* algorithm that makes t cell-probes *in expectation* on a table of size s cells, each of w bits (assuming w+log $s < n$ / log Φ), under the *input distribution*:

database $y=(y_1, y_2,\ldots,y_n)$ where $y_1, y_2,\ldots,y_n \sim \mu$, *i.i.d.*

query $x \sim \mu$, *independently*

$$\Rightarrow \quad t = \Omega\left(\frac{\log \Phi}{\log \frac{sw}{n \log \Psi}}\right)$$

Average-Case Lower Bounds

Hamming space $X = \{0,1\}^d$ database size: n

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

- database: $y_1,\ldots,y_n \in \{0,1\}^d$ *i.i.d. uniform*
- query: uniform and independent $x \in \{0,1\}^d$

Approximate Near-Neighbor (ANN)		Randomized *Exact* Near-Neighbor (*RENN*)
Deterministic	Randomized	
$t = \Omega\left(\frac{d}{\log s}\right)$ [Miltersen *et al.* 1995] [Liu 2004]	$t = \Omega\left(\frac{\log n}{\log \frac{sw}{n}}\right)$ [Panigrahy Talwar Wieder 2008, 2010]	$t = \Omega\left(\frac{d}{\log s}\right)$ [Borodin Ostrovsky Rabani 1999] [Barkol Rabani 2000]
our result: $t = \Omega\left(\frac{d}{\log \frac{sw}{nd}}\right)$		

Thank you!