Class 8

Locality Sensitive Hashing

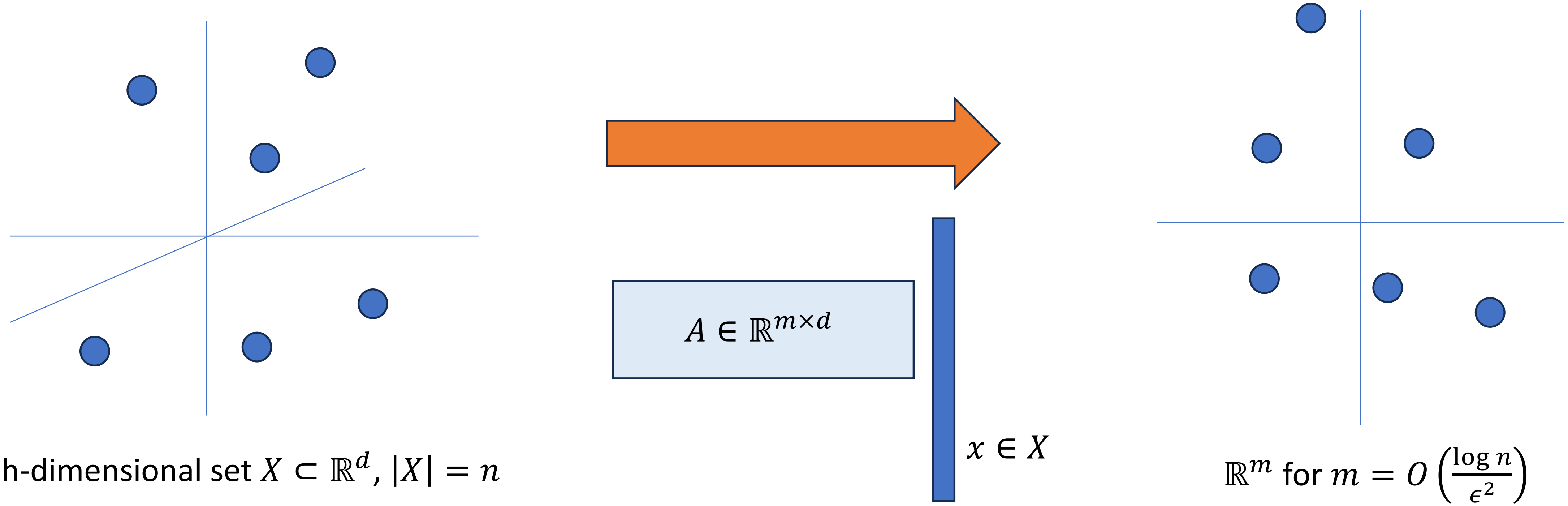
Announcements

- HW3 Due Friday!

Recap: JL!

Lemma: If we choose the entries of A to be i.i.d. $N\left(0, \frac{1}{m}\right)$, then whp this is an embedding of X with distortion at most $(1+\epsilon)$.

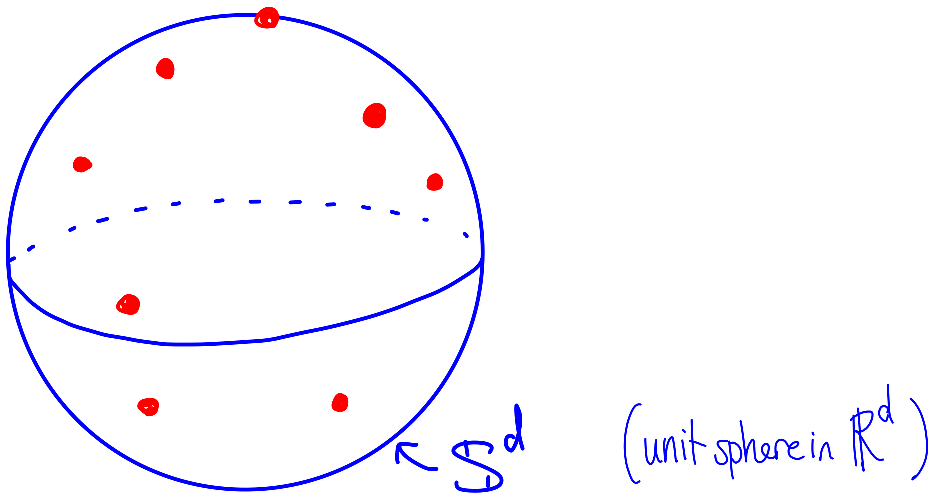
Aka, $\forall\, x, y \in X$, $||Ax - Ay||_2 = (1 \pm \epsilon)||x - y||_2$



High-dimensional set $X \subset \mathbb{R}^d$, $|X| = n$

$\mathbb{R}^m$ for $m = O\left(\frac{\log n}{\epsilon^2}\right)$

Recap: Intro to Nearest Neighbors



$S = \{x_1, x_2, \ldots, x_n\} \subseteq S^d$

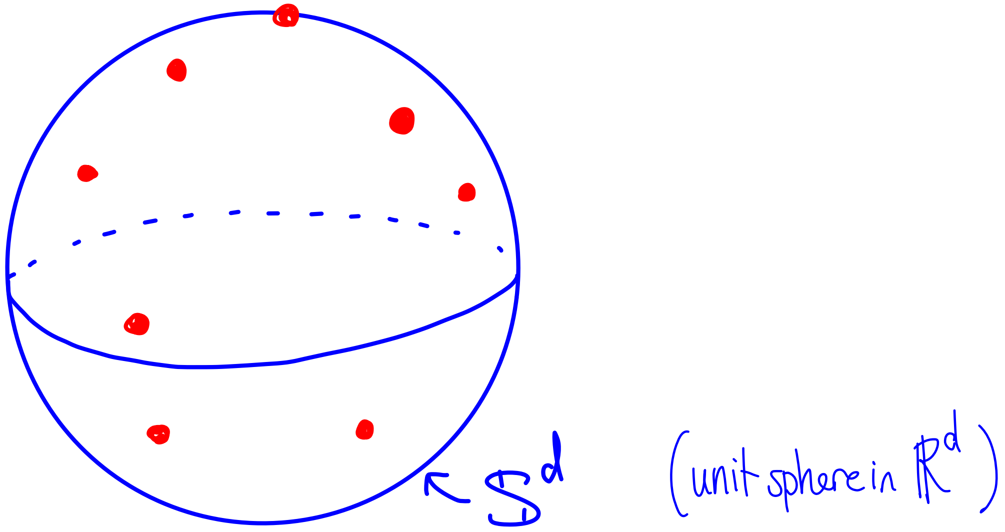
(For today all of our points live on the unit sphere.)

Exact NN problem:

Given $y \in \mathbb{R}^d$, find $\mathrm{argmin}_i ||x_i - y||_2$

Method	Space	Query Time
Linear scan	$O(nd)$ 😃	$O(nd)$ ☹
Various ways of generalizing the $d = 1$ solution	$n^{O(d)}$ ☹	$d^{O(1)} \log n$ 😃
Other heuristics	$O(nd)$ 😃	$\Omega(n)$ in the worst case ☹

Recap: c-near neighbors



$S = \{x_1, x_2, \ldots, x_n\} \subseteq S^d$

(For today all of our points live on the unit sphere.)

Goal: Given y, find x_j so that

$$\left|\left|y - x_j\right|\right|_2 \leq c \left(\min_i \left|\left|y - x_i\right|\right|_2 \right)$$

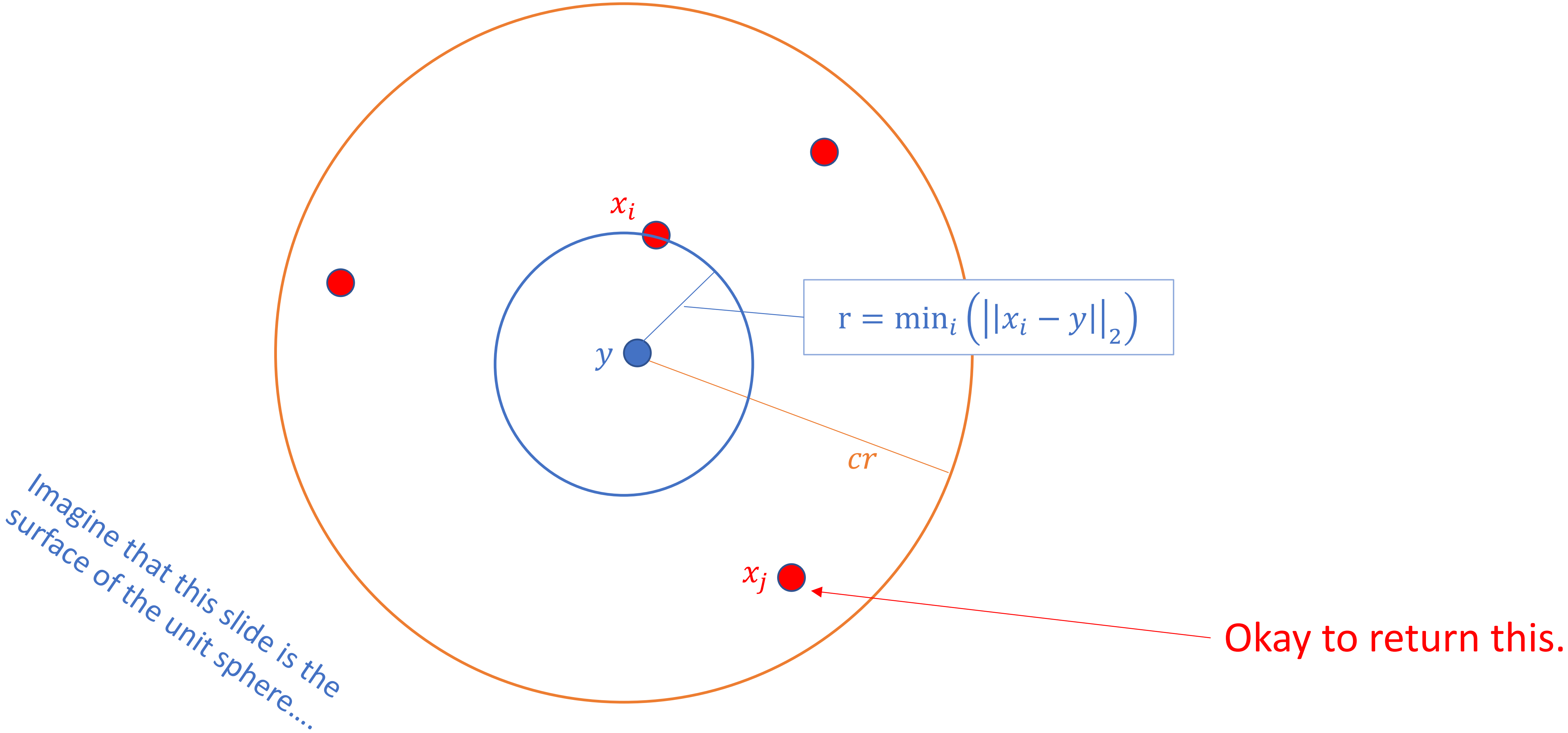
Want:

- Space polynomial in $d \cdot n$
- Query time $o(n)$

Recap: c-near-neighbors

Goal: Given y, find x_j so that

$$\left\|y - x_j\right\|_2 \leq c\left(\min_i \left\|y - x_i\right\|_2\right)$$

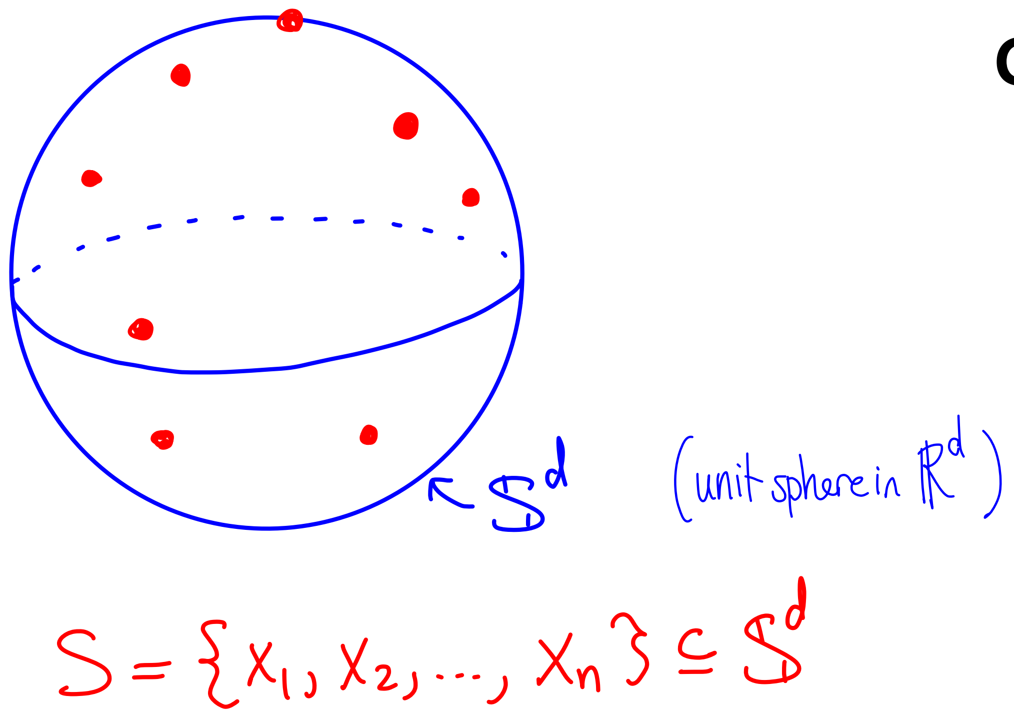


Questions?

Quiz, Mini-Lectures?

Before: Given y, find x_j so that $\left\|y - x_j\right\|_2 \le c\left(\min_i \|y - x_i\|_2\right)$

Today: (r, c)-near-neighbors



(For today all of our points live on the unit sphere.)

Goal: Given y so that $\min_i \left\|y - x_i\right\|_2 \le r$, find x_j so that $\left\|y - x_j\right\|_2 \le cr$

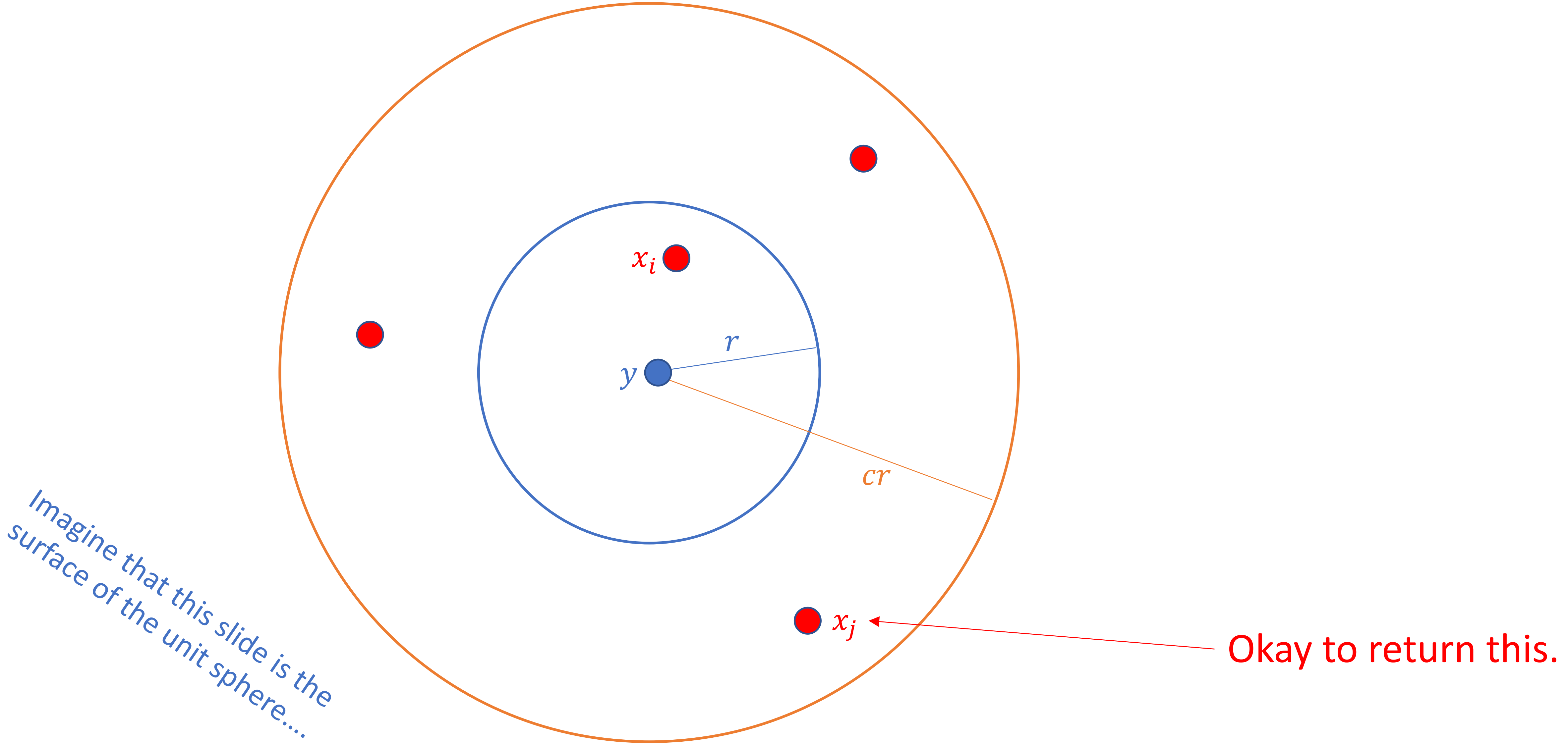
Want:

- Space polynomial in $d \cdot n$
- Query time $o(n)$

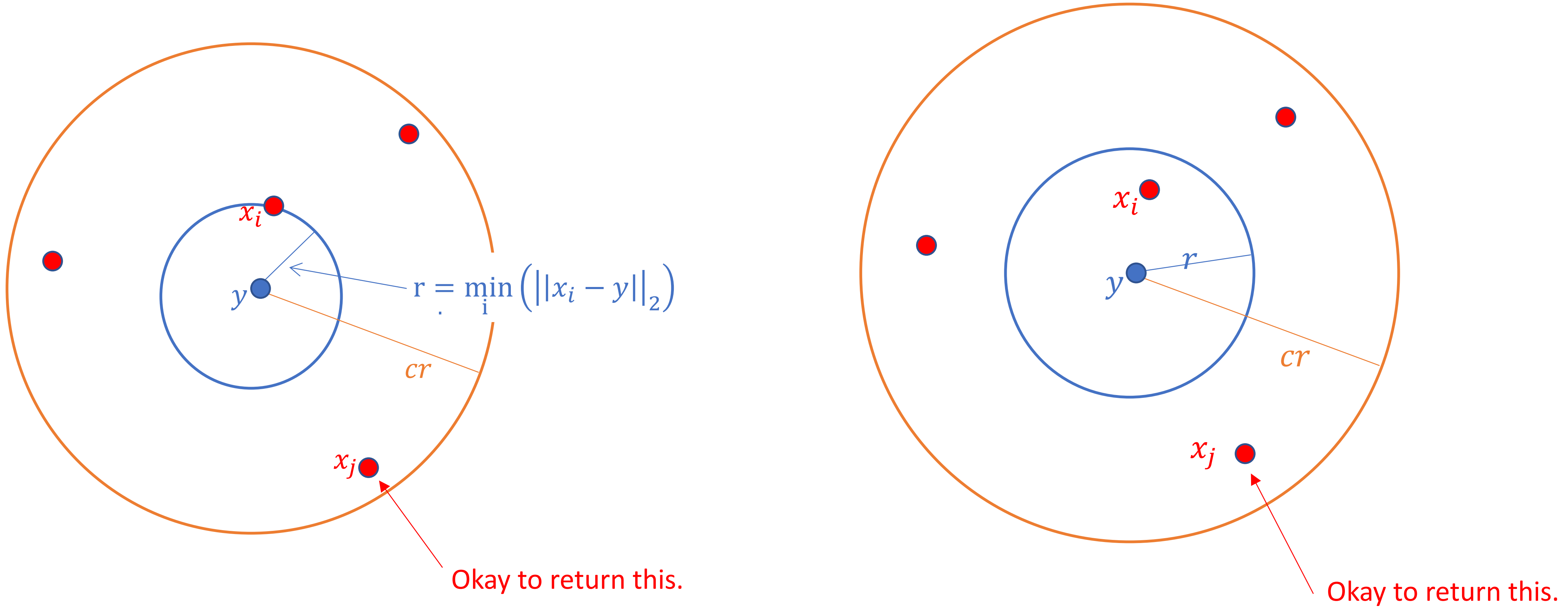
(r, c)-near-neighbors

Goal: Given y so that $\min_i ||y - x_i||_2 \leq r$,

find x_j so that $||y - x_j||_2 \leq cr$



c-NN vs (r, c)-NN


x_i
y
$r = \min_i \left(||x_i - y||_2 \right)$
cr
x_j
Okay to return this.
x_i
r
y
cr
x_j
Okay to return this.

Fact

- If you can solve (r, c)-nearest neighbors then you can (basically) solve c-nearest neighbors.

- (See lecture notes).

Goal for today

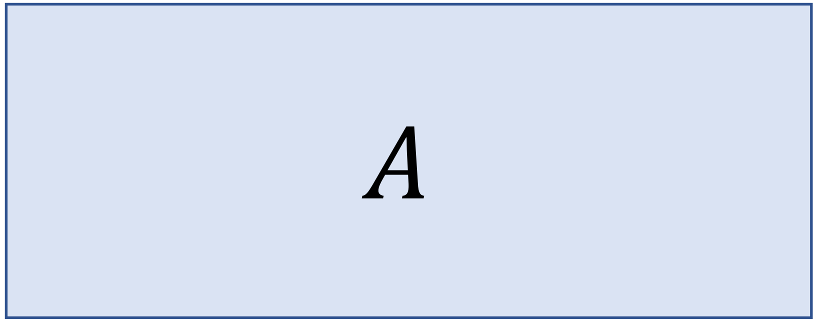
- A solution to (r, c)-approximate nearest neighbors.
- Tool: **Locality-Sensitive Hashing.**
  - Points that are near to each other have a good probability of colliding.
  - Points that are far from each other are unlikely to collide.
- Strategy:
  - Data structure: hash all the x_i
  - To query, hash y. Return anything in y's bucket.

Our strategy will actually be slightly more complicated than this, but this is the basic idea…

Our Locality Sensitive Hashing Scheme

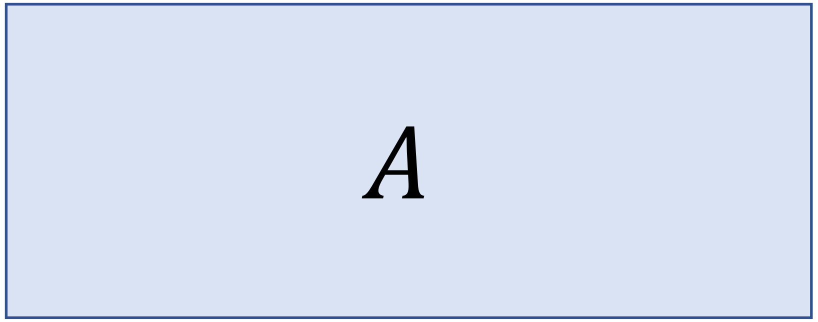
Our Locality Sensitive Hashing Scheme

- Let $A \in \mathbb{R}^{k \times d}$ have i.i.d. $N(0,1)$ entries.
  - Here, $k = \frac{\pi \log n}{2r}$ (we'll see why later).



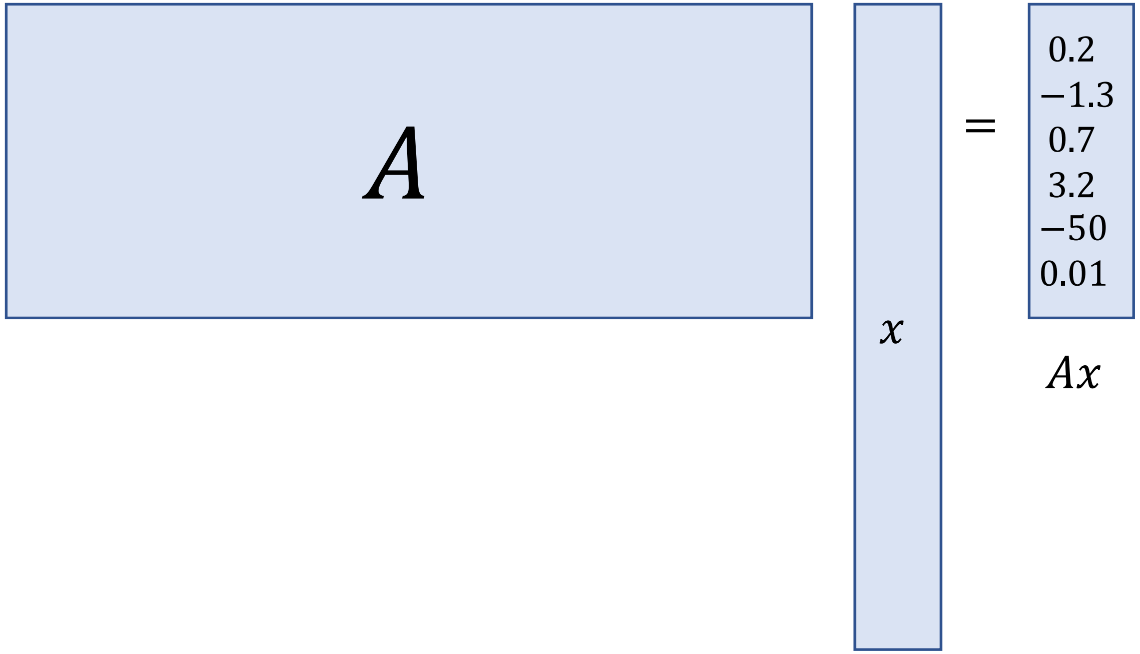
Our Locality Sensitive Hashing Scheme

- Let $A \in \mathbb{R}^{k \times d}$ have i.i.d. $N(0,1)$ entries.
  - Here, $k = \frac{\pi \log n}{2r}$ (we'll see why later).
- Define $h(x) = \text{sign}(Ax)$



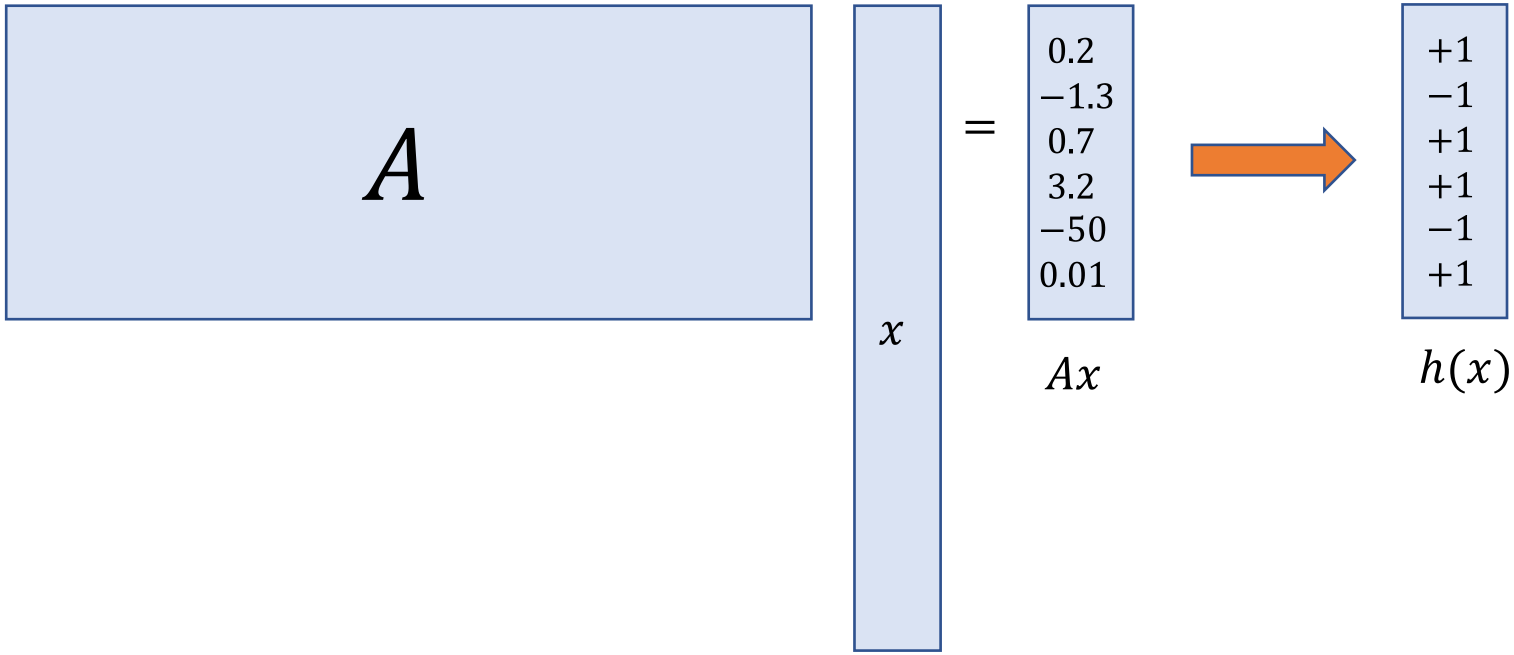
Our Locality Sensitive Hashing Scheme

- Let $A \in \mathbb{R}^{k \times d}$ have i.i.d. $N(0,1)$ entries.
  - Here, $k = \frac{\pi \log n}{2r}$ (we'll see why later).
- Define $h(x) = \text{sign}(Ax)$



Our Locality Sensitive Hashing Scheme

- Let $A \in \mathbb{R}^{k \times d}$ have i.i.d. $N(0,1)$ entries.
  - Here, $k = \frac{\pi \log n}{2r}$ (we'll see why later).
- Define $h(x) = \text{sign}(Ax)$



Actually choose s independent copies of this

- Choose $s = \sqrt{n}$
- Choose $k = \frac{\pi \log n}{2r}$
- For $i = 1, \ldots, s$:
  - Let $A_i \in \mathbb{R}^{k \times d}$ have i.i.d. $N(0,1)$ entries.
  - Define $h_i(x) = \mathrm{sign}(A_i x)$

Outline of group work

- First (problems 1-5) you will show that:
  - If x, y are close, then probably there's some i so that $h_i(x) = h_i(y)$
  - If x, y are far, then probably there's no such i.

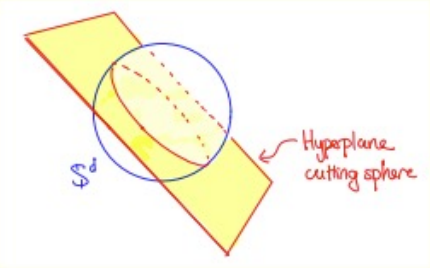
- Then (problems 6,7), you will show how to use this to get a (c, r)-near-neighbors scheme.

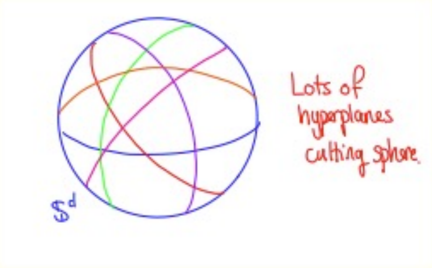
Group work!

(the font is small…look at handouts!)

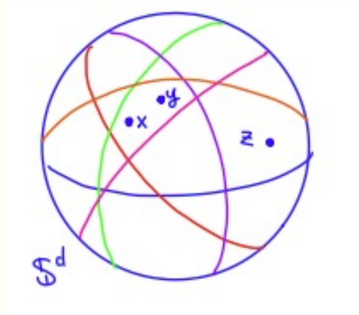
1. Consider a hash function $h_i : \mathbb{S}^d \to \{\pm 1\}^k$ as defined above. Explain why "$h_i(x) = h_i(y)$" has the following geometric meaning:

   Imagine choosing k uniformly random hyperplanes in $\mathbb{R}^d$, and using them to slice up the sphere $\mathbb{S}^d$ like this:





   Then $h_i(x) = h_i(y)$ if and only if x and y are in the same "cell" of this slicing. For example, in the picture below $h_i(x) = h_i(y) \neq h_i(z)$.



2. Explain why, for $x, y \in \mathbb{S}^d$, and for any $i = 1, \ldots, s$,

$$\Pr[h_i(x) = h_i(y)] = \left(1 - \frac{\text{angle}(x,y)}{\pi}\right)^k,$$

   where $\text{angle}(x,y) = \arccos(x^T y)$ is the arc-cosine of the dot product of x and y, aka, the angle between x and y.

   Hint: *Think about the geometric intuition in the plane spanned by x and y.*

3. Suppose that $x, y \in \mathbb{S}^d$. Fill in the blank, using the previous part:

$$\Pr[\forall i \in \{1, \ldots, s\}, h_i(x) \neq h_i(y)] = \_\_\_\_\_$$

   (Don't worry about simplifying, you'll do that in the next part).

4. Let $x, y \in \mathbb{S}^d$ and suppose that the angle between x and y is pretty small. Using our choices of s and k above, along with extremely liberal use of the approximation that $1 - x \approx e^{-x}$ for small x, convince yourself that

$$\Pr[\forall i \in \{1, \ldots, s\}, h_i(x) \neq h_i(y)] \approx \exp\left(-n^{1/2 - \text{angle}(x,y)/(2r)}\right).$$

5. Fill in the blanks (assuming that your approximation from the previous step is valid):

   (a) If $\text{angle}(x,y) \leq r$, then

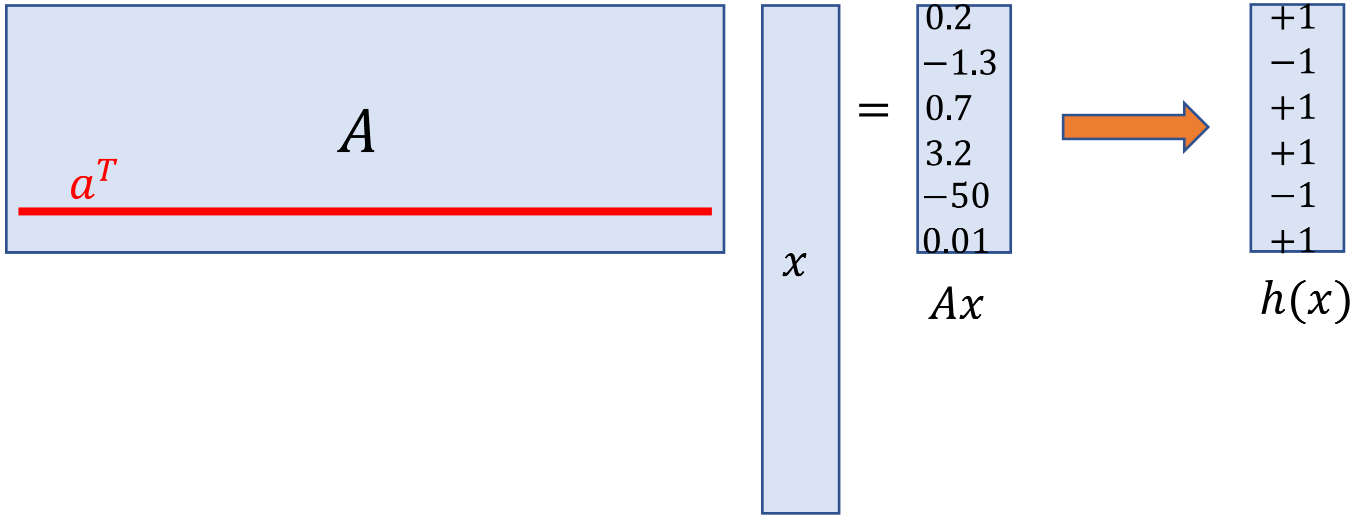
$$\Pr[\exists i \in \{1, \ldots, s\} \text{ so that } h_i(x) = h_i(y)] \geq \_\_\_\_\_\_$$

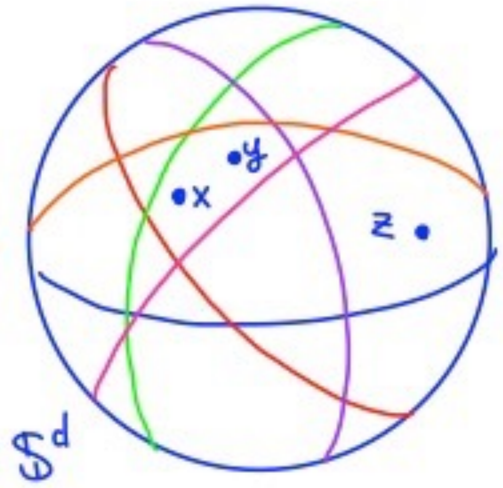
   (b) If $\text{angle}(x,y) \geq 5r$, then

$$\Pr[\exists i \in \{1, \ldots, s\} \text{ so that } h_i(x) = h_i(y)] \leq \_\_\_\_\_\_.$$

Question 1

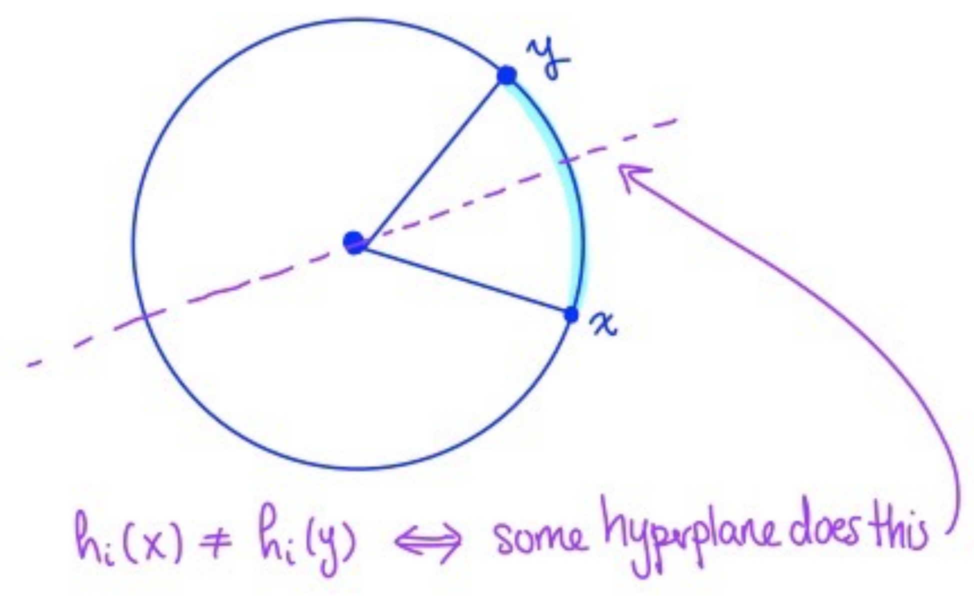
- For each row a^T of A, we have a hyperplane $\{x \in \mathbb{R}^d : a^T x = 0\}$.
- If the corresponding coordinate of Ax is negative, then x lies on one side of the hyperplane, else x lies on the other.
- Same cell = same side of every hyperplane = same sign in every coordinate.





Question 2

$$\Pr[h_i(x) = h_i(x)] = (1 - angle(x, y)/\pi)^k$$

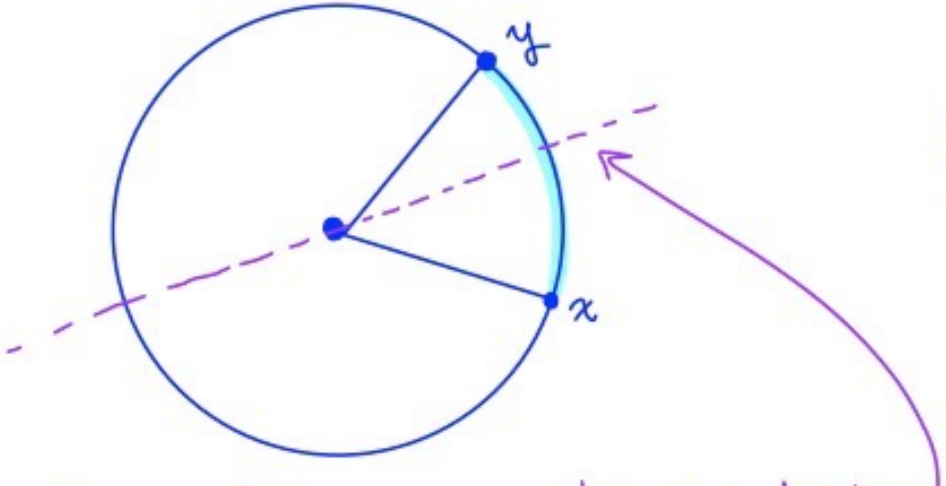


All that matters is what happens in the plane spanned by x and y.

So we can think about a 2-dimensional picture!

Question 2

$$\Pr[h_i(x) = h_i(x)] = (1 - angle(x, y)/\pi)^k$$



$h_i(x) \neq h_i(y) \Leftrightarrow$ some hyperplane does this

$$\mathbb{P}\left\{\begin{array}{l}\text{some hyperplane}\\ \text{does that}\end{array}\right\} = \frac{\text{arclength of arc}}{\pi} = \frac{angle(x,y)}{\pi}$$

$$\mathbb{P}\left\{\begin{array}{l}\text{none of the } k\\ \text{hyperplanes do that}\end{array}\right\} = \left(1 - \frac{angle(x,y)}{\pi}\right)^k$$

Question 3

$$\Pr[\forall i \in \{1, \ldots, s\}, h_i(x) \neq h_i(y)] = \_\_\_\_\_$$

Q2: $\Pr[h_i(x) = h_i(y)] = (1 - angle(x,y)/\pi)^k$

$$\left(1 - \left(1 - \frac{\text{angle}(x,y)}{\pi}\right)^k\right)^s$$

Question 3

$$\Pr[\forall i \in \{1, \ldots, s\}, h_i(x) \neq h_i(y)] = \_\_\_\_\_$$

Q2: $\Pr[h_i(x) = h_i(y)] = (1 - angle(x,y)/\pi)^k$

$$\left(1 - \left(1 - \frac{\text{angle}(x,y)}{\pi}\right)^k\right)^s$$

$\Pr[h_i(x) = h_i(y)]$

Question 3

$$\Pr[\forall i \in \{1, \ldots, s\}, h_i(x) \neq h_i(y)] = \_\_\_\_\_$$

Q2: $\Pr[h_i(x) = h_i(y)] = (1 - angle(x, y)/\pi)^k$

$$\left(1 - \left(1 - \frac{\text{angle}(x, y)}{\pi}\right)^k\right)^s$$

$\Pr[h_i(x) \neq h_i(y)]$

Question 3

$$\Pr[\forall i \in \{1, \ldots, s\}, h_i(x) \neq h_i(y)] = \_\_\_\_\_$$

Q2: $\Pr[h_i(x) = h_i(y)] = (1 - angle(x,y)/\pi)^k$

$$\left(1 - \left(1 - \frac{\text{angle}(x,y)}{\pi}\right)^k\right)^s$$

$\Pr[h_i(x) \neq h_i(y)\ \forall \boldsymbol{i}]$

Question 4

$$\Pr[\, h_i(x) \neq h_i(y) \; \forall \; i \in [s]] \approx \exp\left(-n^{\left(\frac{1}{2} - \frac{\text{angle}(x,y)}{2r}\right)}\right)$$

Assuming $\text{angle}(x, y)$ is pretty small

Recall that $s = \sqrt{n}$, $k = \frac{\pi \log n}{2r}$

$$\Pr[\, h_i(x) \neq h_i(y) \forall i \in [s]] \leq \left(1 - \left(1 - \frac{\text{angle}(x, y)}{\pi}\right)^k\right)^s$$

Question 4

$$\Pr[\, h_i(x) \neq h_i(y)\ \forall\, i \in [s]] \approx \exp\left(-n^{\left(\frac{1}{2} - \frac{\text{angle}(x,y)}{2r}\right)}\right)$$

Assuming $\text{angle}(x, y)$ is pretty small

Recall that $s = \sqrt{n}$, $k = \frac{\pi \log n}{2r}$

$$\Pr[\, h_i(x) \neq h_i(y) \forall i \in [s]] \leq \left(1 - \left(1 - \frac{\text{angle}(x,y)}{\pi}\right)^k\right)^s$$

$$\left(1 - \frac{\text{angle}(x,y)}{\pi}\right)^k \approx \exp\left(-k \cdot \frac{\text{angle}(x,y)}{\pi}\right)$$

$$= \exp\left(\frac{-\log n \cdot \text{angle}(x,y)}{2r}\right) \quad = n^{\frac{-\text{angle}(x,y)}{2r}}$$

Question 4

$$\Pr[\, h_i(x) \neq h_i(y)\ \forall\ i \in [s]] \approx \exp\left(-n^{\left(\frac{1}{2} - \frac{\text{angle}(x,y)}{2r}\right)}\right)$$

Assuming $\text{angle}(x, y)$ is pretty small

Recall that $s = \sqrt{n},\ k = \frac{\pi \log n}{2r}$

$$\Pr[\, h_i(x) \neq h_i(y) \forall i \in [s]] \leq \left(1 - n^{\frac{-\text{angle}(x,y)}{2r}}\right)^s$$

Question 4

$$\Pr[\, h_i(x) \neq h_i(y)\ \forall\, i \in [s]] \approx \exp\left(-n^{\left(\frac{1}{2} - \frac{\text{angle}(x,y)}{2r}\right)}\right)$$

Assuming $\text{angle}(x, y)$ is pretty small

Recall that $s = \sqrt{n},\ k = \frac{\pi \log n}{2r}$

$$\Pr[\, h_i(x) \neq h_i(y) \forall i \in [s]] \leq \left(1 - n^{\frac{-\text{angle}(x,y)}{2r}}\right)^{s}$$

$$\approx \exp\left(-s \cdot n^{\frac{-\text{angle}(x,y)}{2r}}\right)$$

$$= \exp\left(-n^{1/2} \cdot n^{\frac{-\text{angle}(x,y)}{2r}}\right)$$

Question 5(a)

If $\text{angle}(x, y) \leq r$,

$$\Pr[\,\exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \geq \underline{\qquad 1 - \frac{1}{e} \qquad}$$

$$\Pr\left[\, h_i(x) \neq h_i(y) \forall i \in [s]\right] \quad \approx \exp\left(-n^{1/2} \cdot n^{\frac{-\text{angle}(x,y)}{2r}}\right)$$

$$\leq \exp\left(-n^{1/2} \cdot n^{-1/2}\right)$$

$$= \frac{1}{e}$$

Question 5(b)

If $\text{angle}(x, y) \geq 5r$,

$$\Pr[\,\exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \leq \underline{\quad O\left(\frac{1}{n^2}\right) \quad}$$

$$\Pr\big[\, h_i(x) \neq h_i(y) \forall i \in [s]\big] \approx \exp\left(-n^{1/2} \cdot n^{\frac{-\text{angle}(x,y)}{2r}}\right)$$

$$\geq \exp\left(-n^{1/2} \cdot n^{-5/2}\right)$$

$$= \exp(-n^{-2}) \quad \approx 1 - \frac{1}{n^2}$$

Locality-Sensitive Hash (LSH) Families

- Let $\mathcal{H}$ be a family of hash functions $h: \mathbb{R}^d \to D$.
- $\mathcal{H}$ is a LSH family (with parameters R, C) if:
  - If $||x - y||_2 \leq R$, then $\Pr_{h \in \mathcal{H}}[\, h(x) = h(y)]$ is big
  - If $||x - y||_2 \geq C \cdot R$, then $\Pr_{h \in \mathcal{H}}[\, h(x) = h(y)]$ is small.

- We just made one of these! (If we replace $angle(x, y)$ with $||x - y||_2$ ...)

$x \in \mathbb{R}^d$

$h_i(x)$

Next up: Use this to solve (c, r)-Near-Neighbors!

Group Work

If $\text{angle}(x, y) \leq r$,

$$\Pr[\, \exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \geq 1 - \frac{1}{e}$$

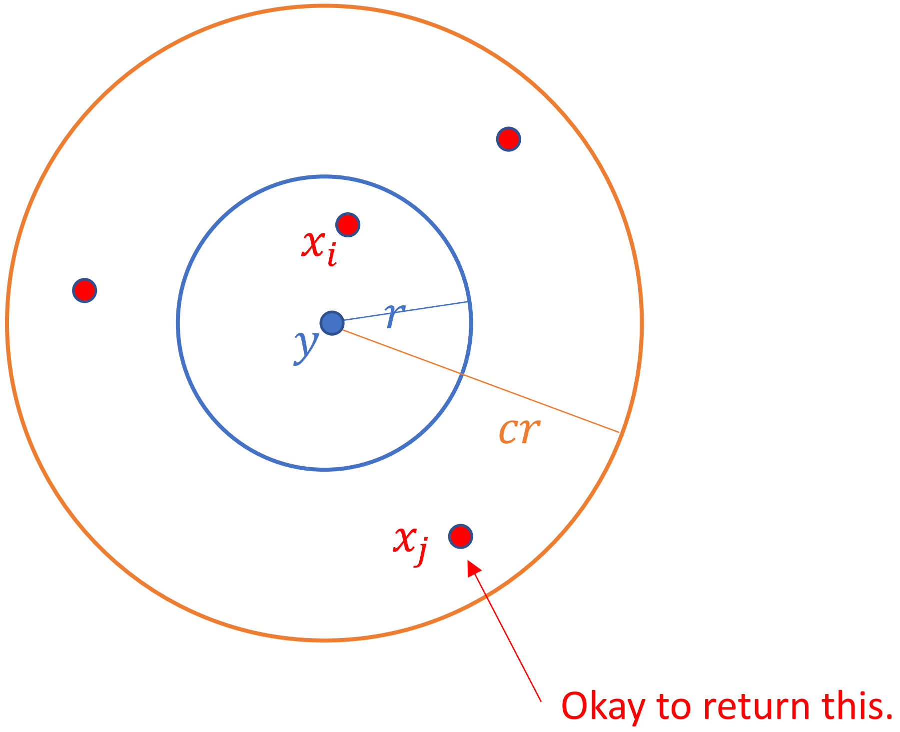
If $\text{angle}(x, y) \geq 5r$,

$$\Pr[\, \exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \leq O\left(\frac{1}{n^2}\right)$$

6. Pretend that angle(x,y) = $||x - y||_2$. Solve (c, r)-NN for some constant c.

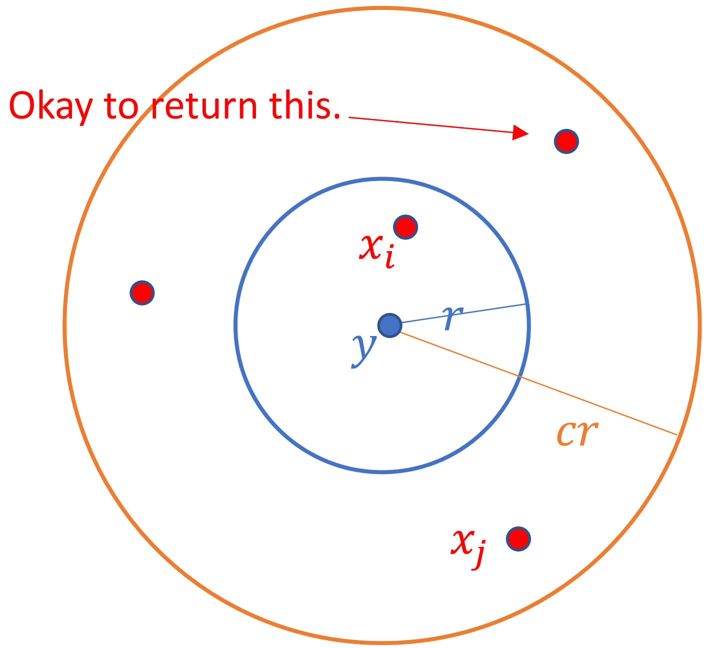
7. Explain why pretending this is okay.

8. How much space does your data structure use? What is the query time?



Question 6

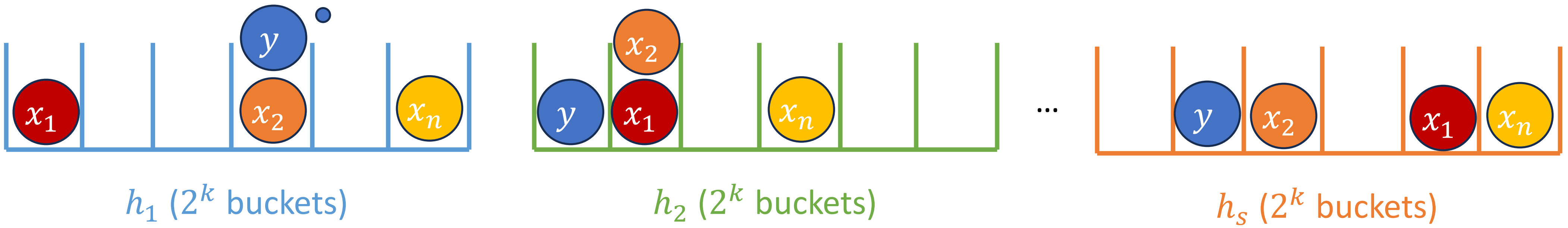
Pretend that angle(x,y) = $||x - y||_2$.
Solve (c, r)-NN for some constant c.



- Data structure:
  - Store s hash tables, one for each i.

- Query Algorithm:
  - For $i = 1,2, \dots, s$:
    - Compute $h_i(y)$
    - If there's some x_j so that $h_i(x_j) = h_i(y)$, return it.

I'm more likely to collide with points I'm close to. Let's say x_2!

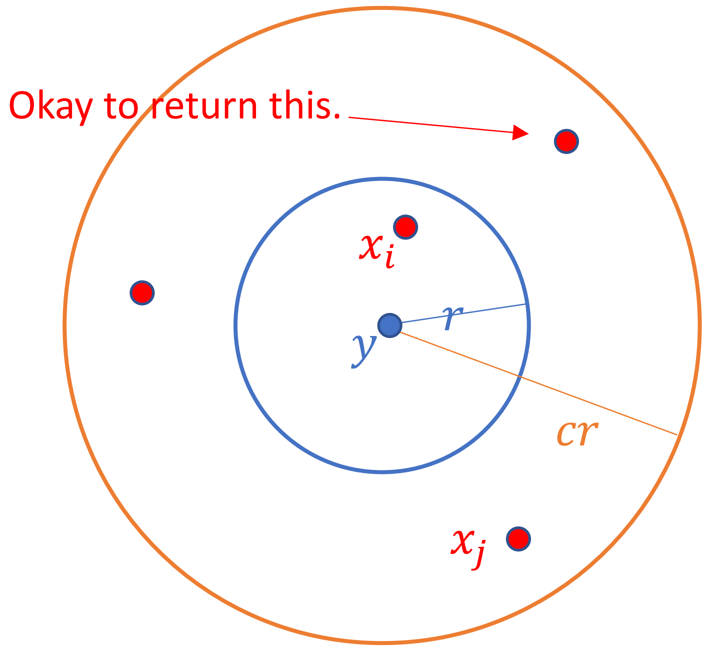


If angle$(x, y) \leq r$, $\Pr[\exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \geq 1 - \frac{1}{e}$

If angle$(x, y) \geq 5r$, $\Pr[\exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \leq O\left(\frac{1}{n^2}\right)$

Question 6

- Query Algorithm:
  - For $i = 1,2, \ldots, s$:
    - Compute $h_i(y)$
    - If there's some x_j so that $h_i(x_j) = h_i(y)$, return it.



- If angle$(y, x_\ell) \leq r$, then with probability at least $1 - \frac{1}{e}$ there's some i so that $h_i(x_\ell) = h_i(y)$.
  - In particular, the algorithm will return **something**.
- If the algorithm returns x_j then with high probability angle$(y, x_j) \leq 5r$.
  - If angle$(y, x_j) > 5r$, $\Pr[\exists i, h_i(x_j) = h_i(y)] \leq \frac{1}{n^2}$, and we can union bound over all x_j to say that never happens whp.

Why is it okay to say "angle(x,y)" is the same as $||x-y||_2$?

FACT: $\frac{2}{\pi}\text{angle}(x,y) \leq ||\text{x}-\text{y}||_2 \leq \text{angle}(x,y)$

Question 7

~~If $\text{angle}(x,y) \leq r$,~~ $\Pr[\exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \geq 1 - \frac{1}{e}$

If $||x-y||_2 \leq \frac{\pi}{2} \cdot r$

~~If $\text{angle}(x,y) \geq 5r$,~~ $\Pr[\exists i \;\; s.t. \;\; h_i(x) = h_i(y)] \leq O\left(\frac{1}{n^2}\right)$

If $||x-y||_2 \geq 5r$

So fiddle around with the value of "c" and the same argument works.

$k = O(\log n)$
$s = \sqrt{n}$

Wrapping up: Time and Space

- Space:

- Update time:

$k = O(\log n)$
$s = \sqrt{n}$

Wrapping up: Time and Space

- Space:
  - s different $k \times d$ matrices A_i: $O(d \cdot \sqrt{n} \cdot \log n)$
  - s hash tables, each with 2^k buckets: $O\left(\sqrt{n} \cdot 2^{O(\log n)}\right) = n^{O(1)}$
  - The elements of S themselves: $O(nd)$
- Update time:

$k = O(\log n)$
$s = \sqrt{n}$

Wrapping up: Time and Space

- Space: $n^{O(1)}$
  - s different $k \times d$ matrices A_i: $O(d \cdot \sqrt{n} \cdot \log n)$
  - s hash tables, each with 2^k buckets: $O\left(\sqrt{n} \cdot 2^{O(\log n)}\right) = n^{O(1)}$
  - The elements of S themselves: $O(nd)$
- Update time:

$k = O(\log n)$
$s = \sqrt{n}$

Wrapping up: Time and Space

- Space: $n^{O(1)}$
  - s different $k \times d$ matrices A_i: $O(d \cdot \sqrt{n} \cdot \log n)$
  - s hash tables, each with 2^k buckets: $O\left(\sqrt{n} \cdot 2^{O(\log n)}\right) = n^{O(1)}$
  - The elements of S themselves: $O(nd)$
- Update time:
  - s different $k \times d$ matrix-vector multiplies: $O(s \cdot k \cdot d) = O(d\sqrt{n} \cdot \log n)$
  - Going through all s hash tables and look in $h_i(y)$'s bucket to see if there's anything else: $O(s) = O(\sqrt{n})$

$k = O(\log n)$
$s = \sqrt{n}$

Wrapping up: Time and Space

- Space: $n^{O(1)}$ 
  - s different $k \times d$ matrices A_i: $O(d \cdot \sqrt{n} \cdot \log n)$
  - s hash tables, each with 2^k buckets: $O\left(\sqrt{n} \cdot 2^{O(\log n)}\right) = n^{O(1)}$
  - The elements of S themselves: $O(nd)$
- Update time: $O(d \cdot \sqrt{n} \cdot \log n) = o(n)$ when d isn't too big. 
  - s different $k \times d$ matrix-vector multiplies: $O(s \cdot k \cdot d) = O(d\sqrt{n} \cdot \log n)$
  - Going through all s hash tables and look in $h_i(y)$'s bucket to see if there's anything else: $O(s) = O(\sqrt{n})$

Recap

- We can use dimension reduction (that smells a bit like JL) to make an efficient c-near-neighbors algorithm!
  - Space $poly(n)$
  - Sub-linear query time!

Next time

- Sparsity!