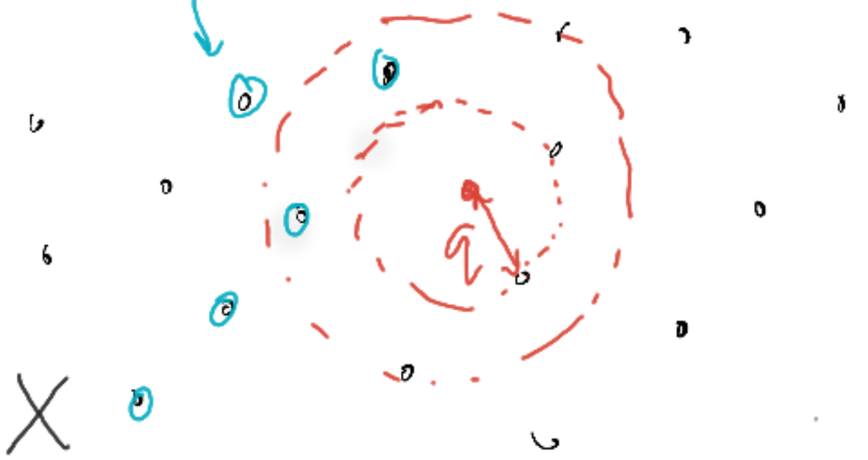
Lecture 10: Locality Sensitive Hashing

- Near(~~est~~) neighbor search
- giving structure
- conventional approaches - $\boxed{\text{kd tree}}$
  - exponential dependence on dimension k

- hashing to the rescue

- build a data structure S
- query q arrives
  return $a \in S$
  $d(q,a) \simeq \min_{a' \in S} d(q,a')$

. Metric space (X, d) ← distance
  $S \subseteq X$ corpus



~~the~~ car was jaguar and

- S = set of strings

each

- document = bag of words

a b c b d a

a	b	c	e	
2	2	1	0	$\boxed{\cdot}$ . . .

#words in doc

$$\text{TF/IDF} \quad w = \frac{\text{freq}(w)}{\text{in my doc}} \times \log\left(\frac{\#\text{words in corpus}}{\#w \text{ appears in corpus}}\right)$$

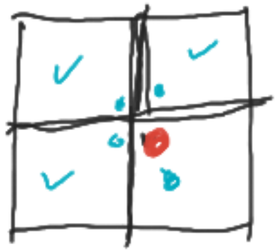
$\mathbb{X} = \mathbb{R}^k$

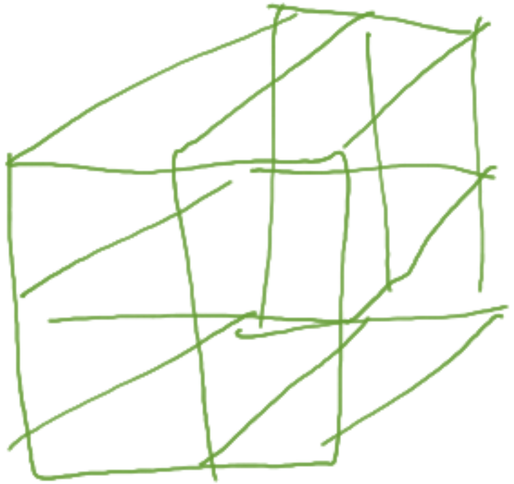
$$d = \|x-y\|_2 = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

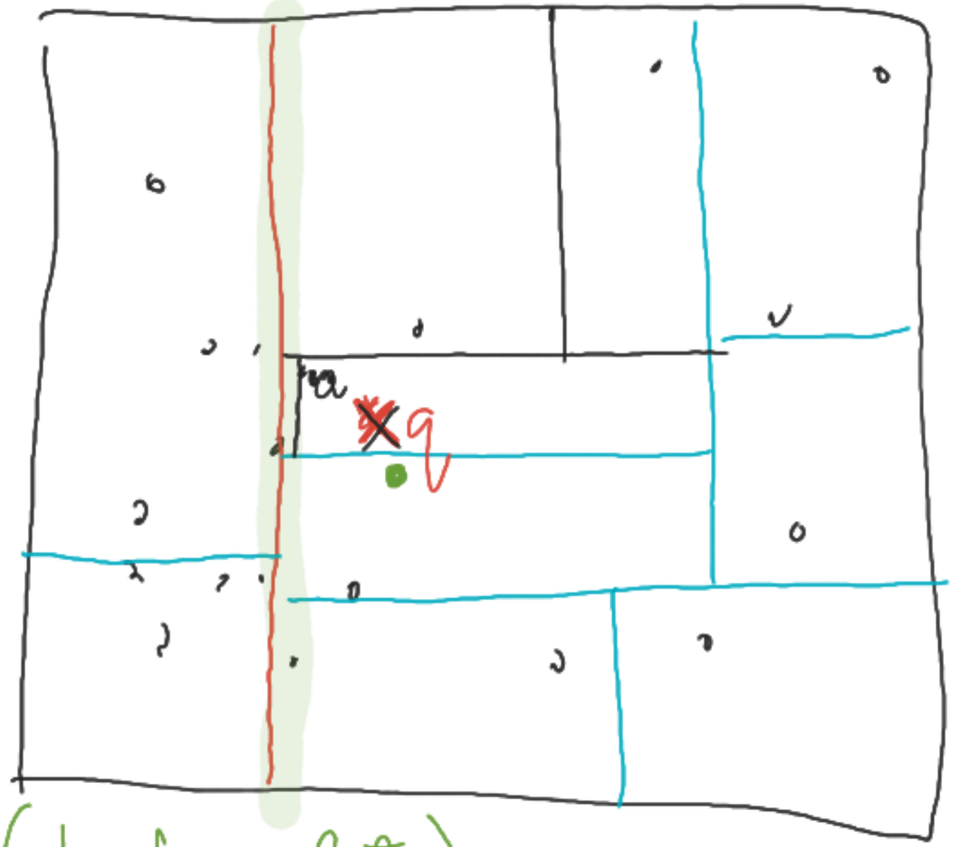
or

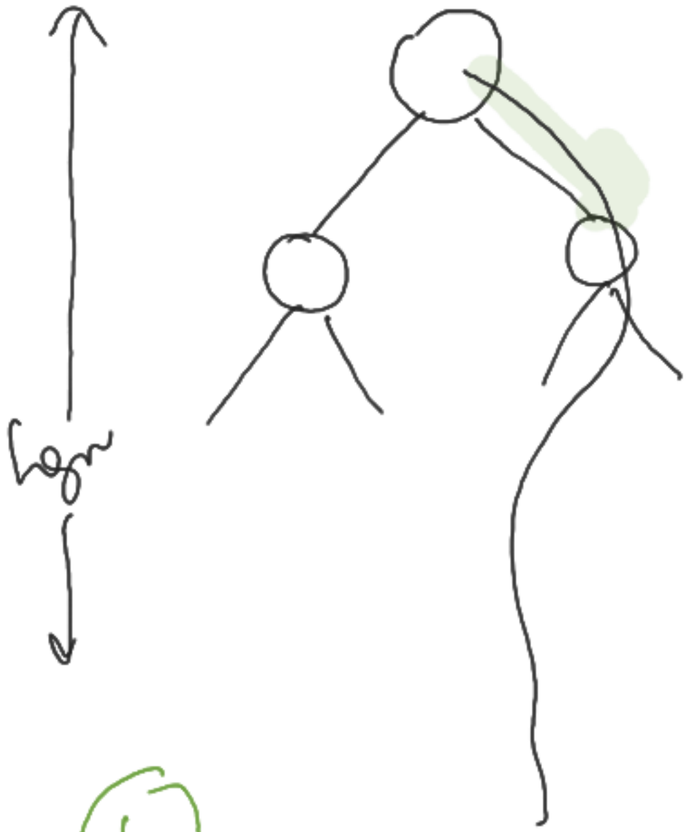
$$\|x-y\|_1 = \sum_i |x_i - y_i|$$

k-d tree









Solutions

1. generic dim red^n (lectures later)
2. LSH

2^k runtime

curse of dimensionality

LSH vs Hashing

$S \subseteq U$ $\quad h : U \to [M]$

if $x = y$ $\quad h(x) = h(y)$

$x \neq y$ $\quad \Pr[h(x) \neq h(y)]$ is high

---

distance between points

x, y close $\quad \Pr[h(x) = h(y)]$ is high

x, y far $\quad \Pr[h(x) \neq h(y)]$ is high

Solving Near Neighbor

$\boxed{\varepsilon\text{-NN}}$

(X,d) metric space $\quad S \subseteq X \quad |S| = n$

Build a data structure

- on seeing queries $q \in X$

  return $a \in S$

  st $d(q,a) \le (1+\varepsilon) \underbrace{\min_{a' \in S} d(q,a')}$

- small space / preprocessing
- small query time

$\boxed{(\varepsilon,r)\text{-NN}}$

Given $\varepsilon, r \in \mathbb{R}_+$

Build a d. structure st $\forall$ query $q \in X$

if $\exists$ a pt $a \in S$ $\quad d(q,a) \le r$

return a pt a'

$d(q,a') \le (1+\varepsilon) r$

if $\exists$ no point at distance $\le r$ from q,

then either return a' st $d(q,a') \le (1+\varepsilon) r$

or return NO point.

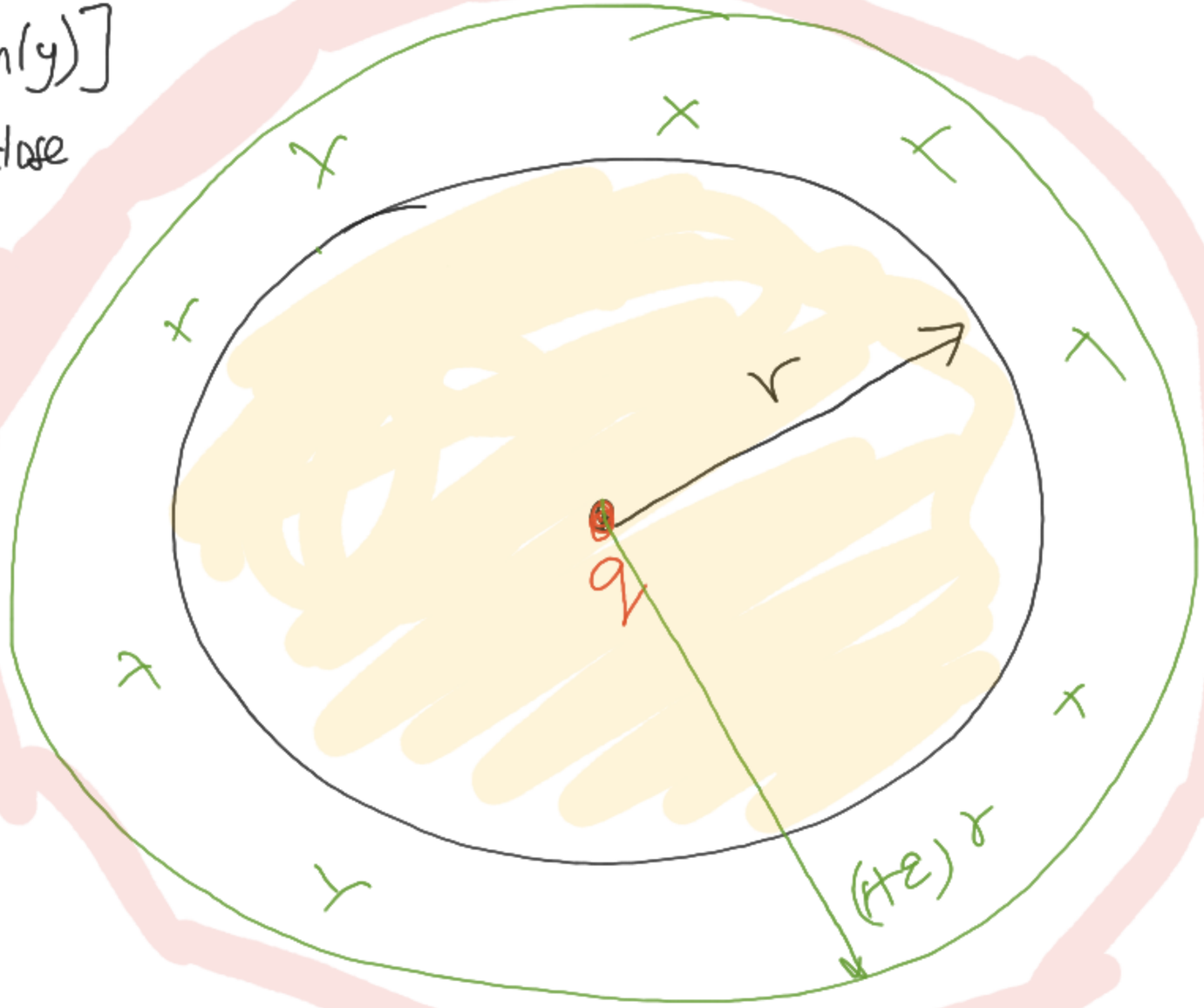
LSH:

$$d(x,y) \le r \Rightarrow \Pr_{h \leftarrow H}[h(x)=h(y)] \geq P_{close}$$

$$d(x,y) \geq (1+\varepsilon)r$$

$$\Rightarrow \Pr_{h \leftarrow H}[h(x)=h(y)] \leq P_{far}$$

$\boxed{H \text{ hash family}}$



LSH for Hamming metric

$X = \{0,1\}^k$ $\qquad$ $d(x,y) = \#$ positions where they differ.

$\overbrace{0\boxed{1}00010111 0}$

① Weak LSH:

metric dependent

Pick a random bit position i of x,

$h(x) = x_i$

$\Pr[h(x) = h(y)] = 1 - \frac{d(x,y)}{k}$

if x, y close $\quad d(x,y) \le r$ $\quad \Rightarrow \Pr(\text{collision}) \ge 1 - \frac{r}{k} \overset{?!}{\approx} 1$

far $\quad \ge 2r$ $\quad \le 1 - \frac{2r}{k} \overset{??}{\approx} 0$

$\varepsilon = 1$

Step 2: Parallel Repetition H $\boxed{\Pr(\text{collision}) \le p}$

H' $h'(x) = (\underbrace{h_1(x), h_2(x) \quad h_t(x)}_{\text{indep}})$

$$\Pr[h'(x) = h'(y)] \le p^t$$

x, y close $\Pr(\text{collision under } h') \ge \left(1 - \frac{r}{K}\right)^t \overset{!!}{\cong} e^{-\frac{r}{K} t} \cong \frac{1}{\sqrt{n}}$

x, y far $\Pr(\text{collision under } h') \le \left(1 - \frac{2r}{K}\right)^t \le e^{-\frac{2r}{K} \cdot t} \overset{\approx 0}{\not\approx} \frac{1}{n}$

☺ if $t = \frac{K}{2r} \cdot \log n$

$\boxed{1 + x \le e^x \ \forall x}$