

LECTURE 11

METRIC SPACE (a set with a distance)

$$(X, d: X \times X \rightarrow \mathbb{R}^+)$$

$$\begin{aligned}d(u, u) &= 0 \\d(u, v) &= d(v, u)\end{aligned}$$

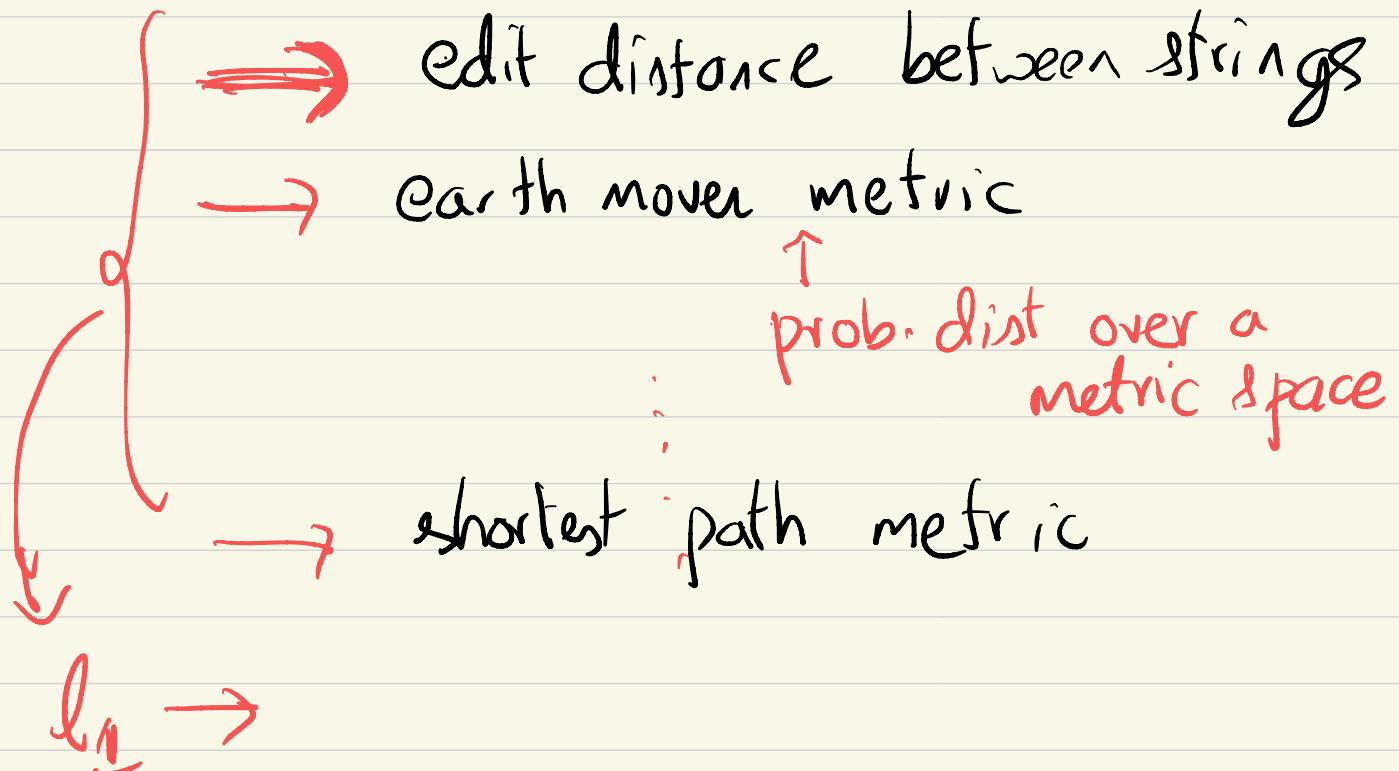
[Triangle Inequality] $d(u, v) + d(v, w) \geq d(u, w)$

Examples:

Euclidean Space

$$\begin{matrix} l_2^d \\ l_1^d \end{matrix}$$

L_1 metric



$D \geq 1$

Embedding: A function $\phi: (X, d) \rightarrow (X', d')$

is a distortion - D embedding if

$\forall x, y \in (X, d)$

$$\frac{d(u, y)}{D} \leq d(\phi(u), \phi(y)) \leq d(x, y)$$

D

Thm: For Any $P \subseteq (\mathbb{R}^d, \|\cdot\|_2)$ $|P|=n$

$\exists \phi$



$\ell_2^{(\log n / \epsilon^2)}$ with distortion $(1+\epsilon)$

Nearest neighbor problem

$(\mathbb{R}^d, \|\cdot\|_2)$

INPUT: A set of n points $P \subseteq (\mathbb{X}, d)$
(Preprocess data) ??
Query $q \in \mathbb{X}$

GOAL: Find the nearest point to
 q in P .

Naive Brute force: Space: $\Theta(nd)$
Query time: $\Theta(nd)$

$(\mathbb{R}, \|\cdot\|_2)$ \rightarrow sort & binary search

$(\mathbb{R}^d, \|\cdot\|_2)$ \rightarrow { K-D tree
Voronoi diagrams }

Space needed grows \Rightarrow expld)

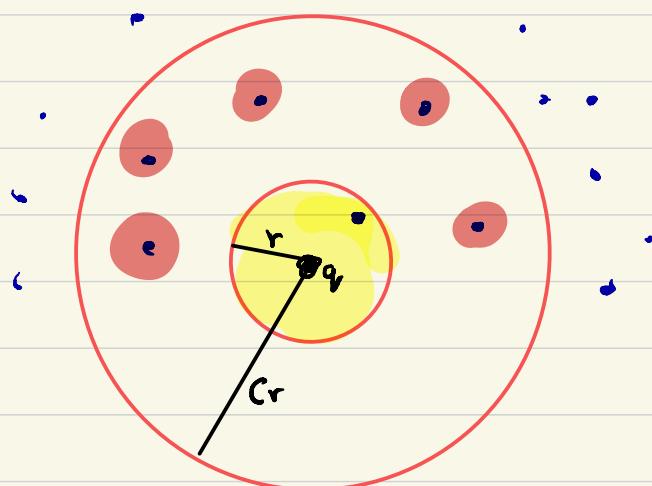
JL Dim reduction : $\log n / \varepsilon^2$

$$\Rightarrow \text{Space: } 2^{\log n / \varepsilon^2} \approx \boxed{n^{1/\varepsilon^2}}$$

Defn: "c-approximate $\xrightarrow{r - \text{near}}$ $\xrightarrow{\text{neighbor}}$ " (ANN1)

Input: $P \subseteq (X, d)$ $|P| = n$

Query: $q \in X$



If \exists some point p
 $d(p, q) \leq r$
then algo must
output p'
 $d(p', q) \leq cr$

c-approximate Nearest Neighbor (ANN2)

Input: $P \subseteq (X, d)$

Query: $q \in X$

Algo outputs p' s.t

$$d(p', q) \leq c \cdot \min_{p \in P} d(p, q)$$

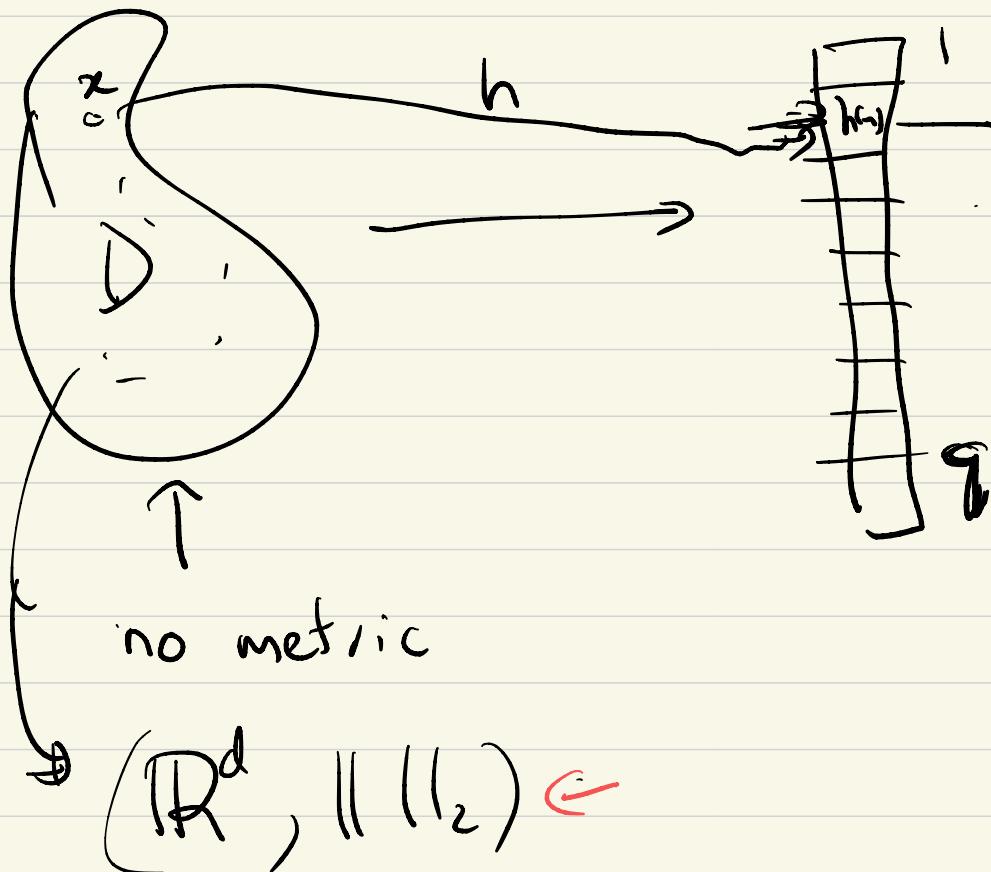
LOCALITY SENSITIVE HASHING

(Universal Hash Family)

$$\mathcal{H} = \{ h : \mathcal{D} \rightarrow \{1..q\} \} \text{ such that}$$

$\forall x \neq y \in \mathcal{D}$

$$\Pr_{h \sim \mathcal{H}} [h(x) \neq h(y)] = 1 - \frac{1}{q} = O(1)$$



Locality Sensitive Hashing

$$f = \frac{\log P_{close}}{\log P_{far}}$$

A $\mathcal{H} = \{ h : (X, d) \rightarrow \mathbb{Z} \}$ is

a $(r, C, P_{close}, P_{far})$ -LSH

$\forall x, y \in (X, d)$

$$d(x, y) \leq r \Rightarrow \Pr_h [h(x) = h(y)] \geq P_{close}$$

$$d(x, y) \geq C \cdot r \Rightarrow \Pr_h [h(x) = h(y)] \leq P_{far}$$

Amplifying the gap

$$G = \underbrace{\{ g = (h_1, h_2, \dots, h_t) : (X, d) \rightarrow \mathbb{Z}^t \}}$$

$\forall x, y$

$$d(x, y) \leq r \Rightarrow \Pr [g(x) = g(y)] \geq P_{close}^{(0.6)^t}$$

$$d(x, y) \geq C_r \Rightarrow \Pr [g(x) = g(y)] \leq P_{far}^{(0.4)^t}$$

For every point $x \approx$ expect about $\leq O(1)$ other points to collide

$$\Pr[g(y) = g(x)] \approx \frac{1}{n}$$

$$\Rightarrow t = \log n / \log(\gamma_{P_{\text{far}}})$$

then for any $d(y, x) \geq Cr$

$$\Pr[g(x) = g(y)] \leq \frac{1}{n}$$

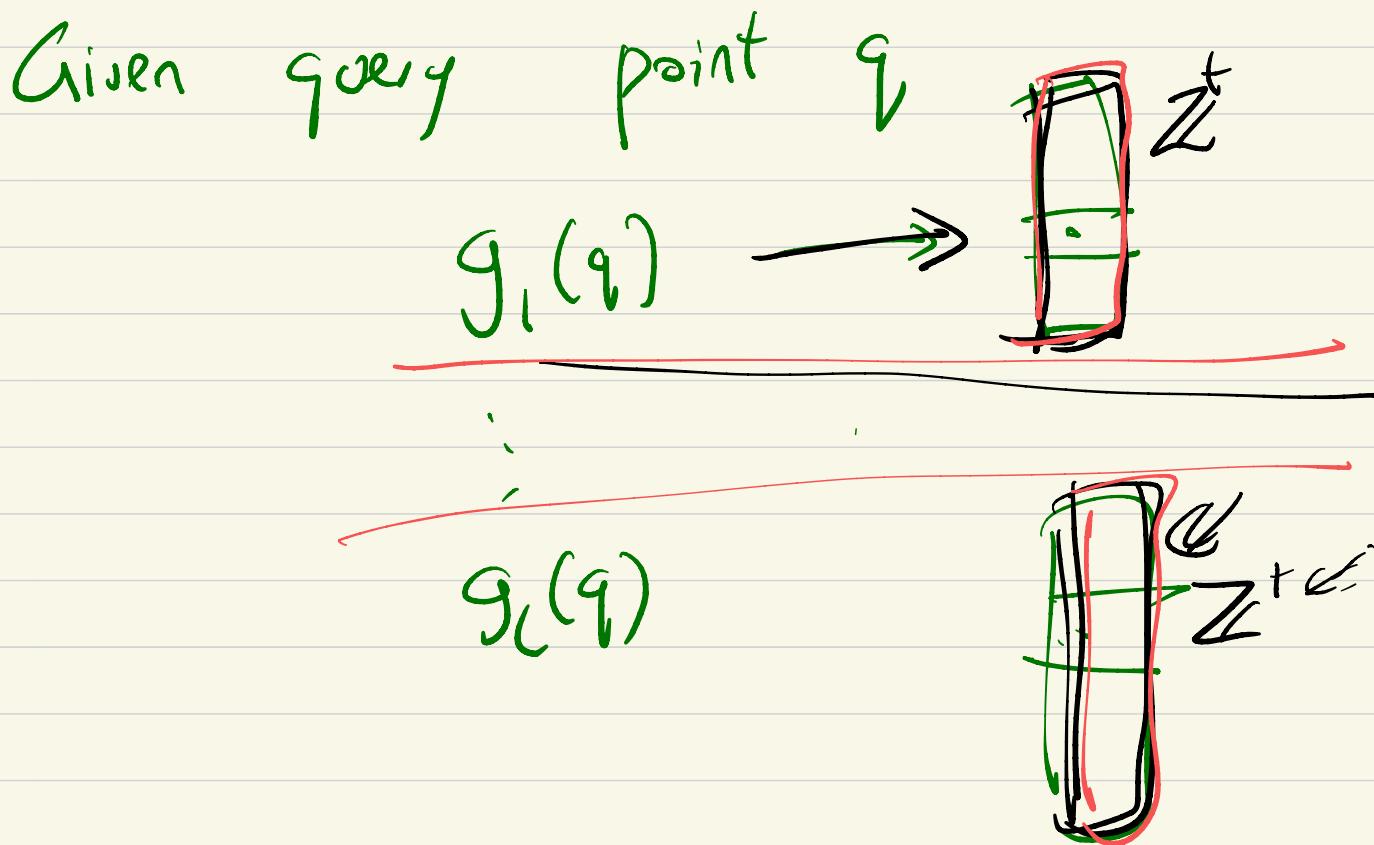
Suppose $d(x, y) < r$

$$\Pr[g(x) = g(y)] \geq (P_{\text{close}})^t \approx \frac{1}{n^\beta}$$

$$\text{where } P_{\text{close}} = P_{\text{far}} / \beta = \frac{\log P_{\text{close}}}{\log P_{\text{far}}}$$

$L = n^{\beta} \cdot \underline{\text{hash tables}} \quad \underline{\{g_1, \dots, g_L\}}$

each $g_i = (\underbrace{h_{i1}, \dots, h_{it}}_{\leftarrow t\text{-copies}})$



Storage: n^{β} hash functions

$n^{1+\beta}$ bits of memory

Query: $\rightarrow t \cdot n^{\beta} \cdot$ (distance computation)

$\rightarrow (X = \{0,1\}^d, \text{ Hamming distance})$

$\|x-y\| = \# \text{ of differing bits}$

$$\mathcal{H} = \left\{ h(u) = x_i \quad \left| \begin{array}{l} i = 1 \dots d \\ h: \{0,1\}^d \rightarrow \{0,1\} \end{array} \right. \right\}$$

$$\Pr_i [x_i = y_i] = 1 - \frac{\|x-y\|}{d}$$

$$P_{\text{close}} = 1 - r/d \approx e^{-r/d} \quad \left| \begin{array}{l} \text{for small } r \end{array} \right.$$

$$P_{\text{far}} = 1 - C_r/d \approx e^{-Cr/d}$$

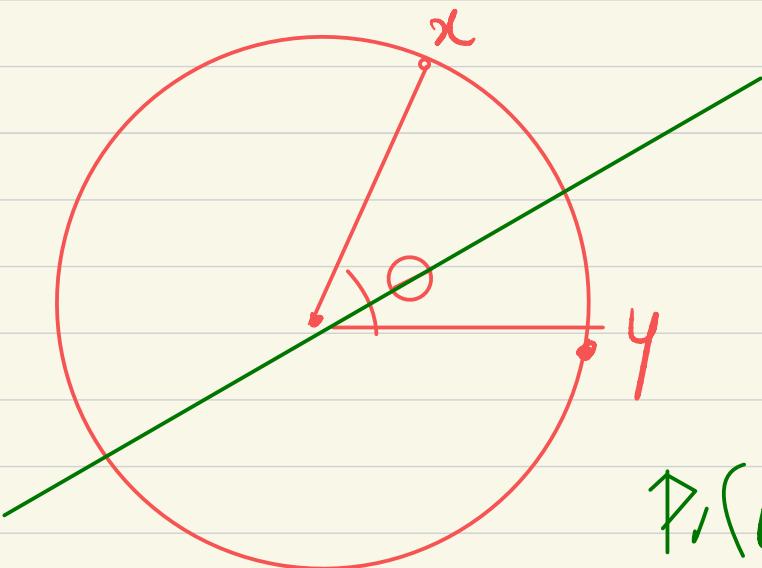
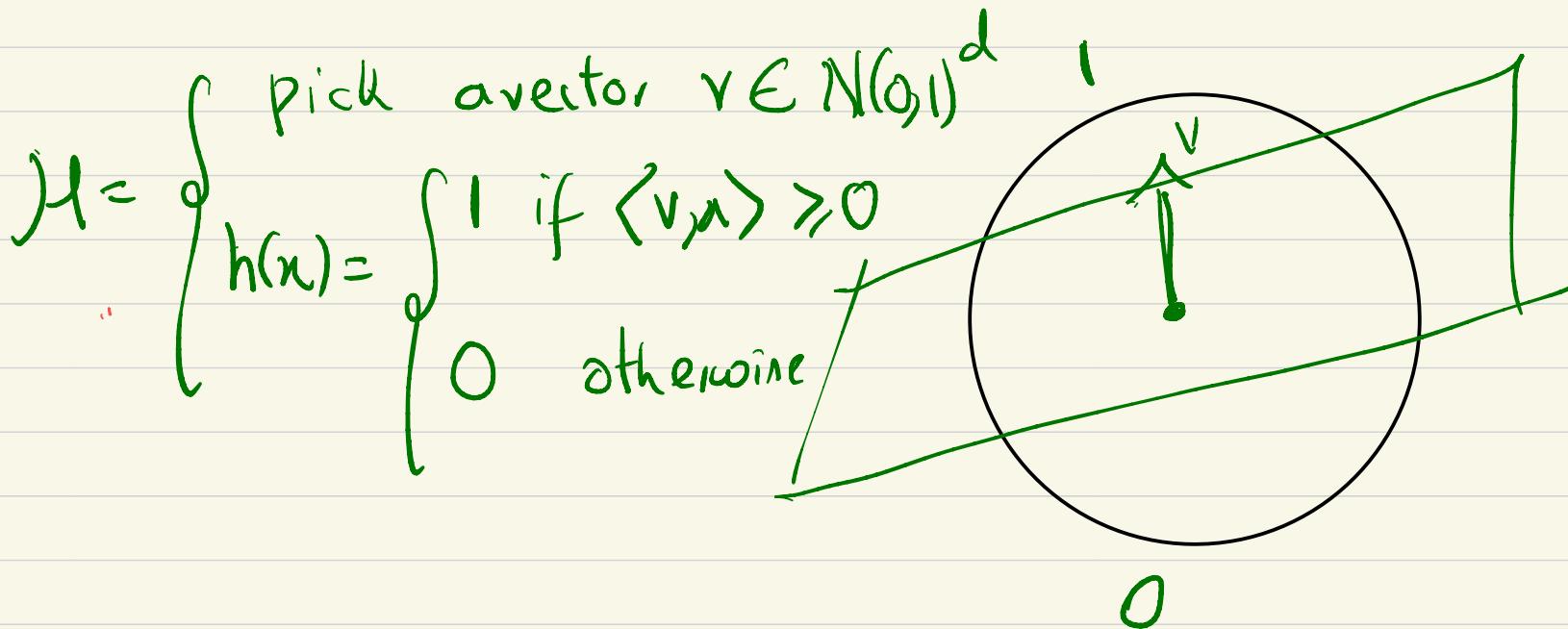
$$\boxed{P = \frac{1}{C}}$$

\Rightarrow space $n^{1+1/C}$
query: $n^{1/C}$

Cosine distance

$$\beta = \frac{1}{C}$$

(unit sphere in \mathbb{R}^d , $d(u, v) = \frac{\delta(u, v)}{\pi}$)



$$\begin{aligned} \Pr(h(x) \neq h(y)) &= \Pr(\text{hyperplane goes between } x \text{ and } y) \\ &= \theta/\pi = d(x, y) \end{aligned}$$

Jaccard Similarity

$$B = \frac{|A \cap B|}{|A \cup B|}$$

↓
Document = { bag of words }
set of words

Given $A, B \subseteq \text{Dictionary}$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \approx 1 \text{ if } A = B \\ \approx 0 \text{ if } A \neq B$$

$$\text{distance} = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \Delta B|}{|A \cup B|}$$

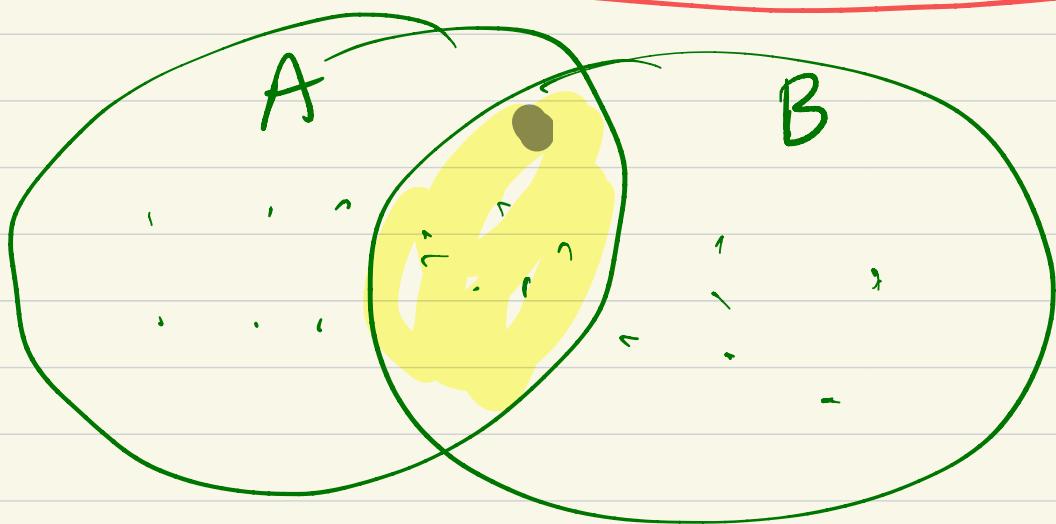
MinHash: [1) Pick a random permutation
 $\pi: D \rightarrow D$]

2) $H_\pi(A) = \text{first word in } A \text{ as per permutation } \pi$.

$A, B \subseteq D$

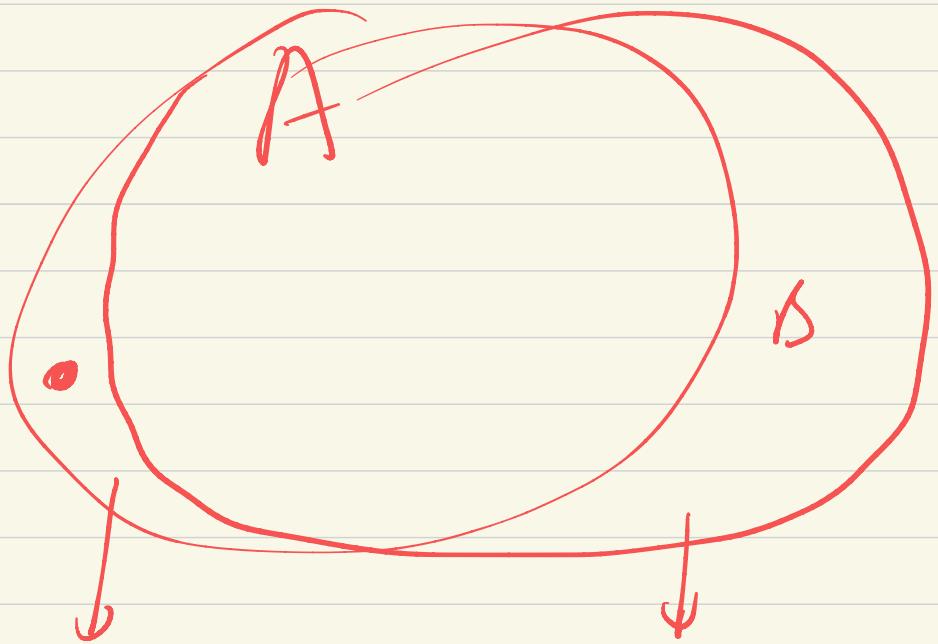
$$\Pr[h_{\pi}(A) = h_{\pi}(B)] = 1 - \text{Jaccard dist}(A, B)$$

$$\Pr[h_{\pi}(A) = h_{\pi}(B)] = \frac{|A \cap B|}{|A \cup B|}$$

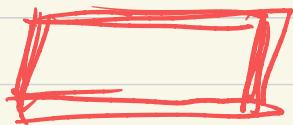


random permutation of all words in $A \cup B$

w.p. $\frac{|A \cap B|}{|A \cup B|}$ the smallest word
in $A \cup B$ falls
in $A \cap B$



Shak



ℓ_2 -distance $(\mathbb{R}^d, \|\cdot\|_2)$

$$(x, y) \in \mathbb{R}^d \quad \text{x}$$

random
projection

1-dimension

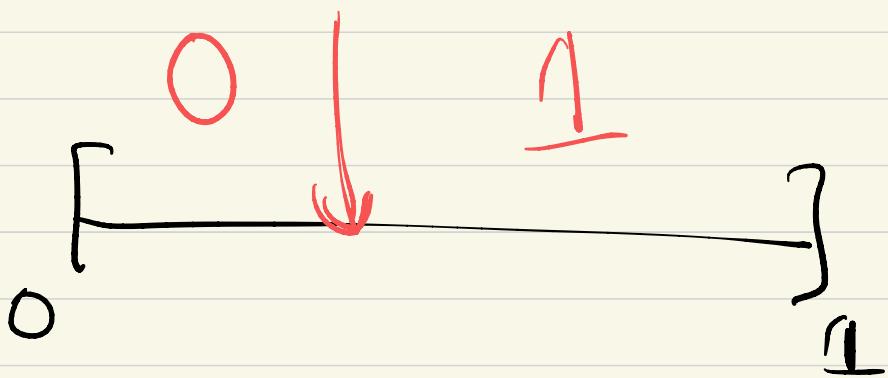
$$a \sim N(0, 1)^d$$

$$([0, 1], \|\cdot\|) \xrightarrow{\text{random projection}} \langle v, x \rangle \in (\mathbb{R})$$

$$\mathbb{E}[\langle v, x \rangle - \langle v, y \rangle] =: \|x - y\|_2$$

$$\{0, 1\}$$

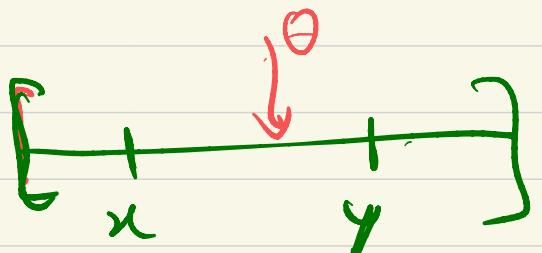
$$\Pr[h(x) \neq h(y)] \propto \|x - y\|$$



Pick a random
 $\theta \in [0, 1]$

$$h(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta \end{cases}$$

$$\Pr[h(u) \neq h(y)]$$



$$= \Pr[\theta \in (x, y)]$$

$$= \frac{|y-x|}{1}$$