

Plan

Logistics

Review

Similarity Estimation

Thanks for coming!

Tea time @ 2 6th Floor

JL Lemma

$$x_1, \dots, x_n \in \mathbb{R}^d \quad k = O\left(\frac{\log n}{\epsilon^2}\right)$$

$\Pi \in \mathbb{R}^{K \times d}$ $\Pi_{i,j}$ = random variable

$$(1-\epsilon) \|x_i - x_j\|_2^2$$

$$\leq \|\Pi x_i - \Pi x_j\|_2^2$$

$$\leq (1+\epsilon) \|x_i - x_j\|_2^2 \quad \text{wp } 9/10$$

Distributional JL

$$k = O\left(\frac{\log(\|S\|)}{\epsilon^2}\right)$$

$$(1-\epsilon) \|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\epsilon) \|x\|_2^2$$

\uparrow
Proved with clever tricks :: wp 1-8

What about inner product?

$$|\langle x_i, x_j \rangle - \langle \Pi x_i, \Pi x_j \rangle| \leq \frac{\epsilon}{2} (\|x_i\|_2^2 + \|x_j\|_2^2)$$

Using JL Lemma!

Application: Fast Set Join estimation

x = people in class

y = people who climb

$x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \leftarrow$ Sujay $y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \leftarrow$ Iris
 \downarrow Iris \downarrow Aidan

$$\langle x, y \rangle = |x \cap y|$$

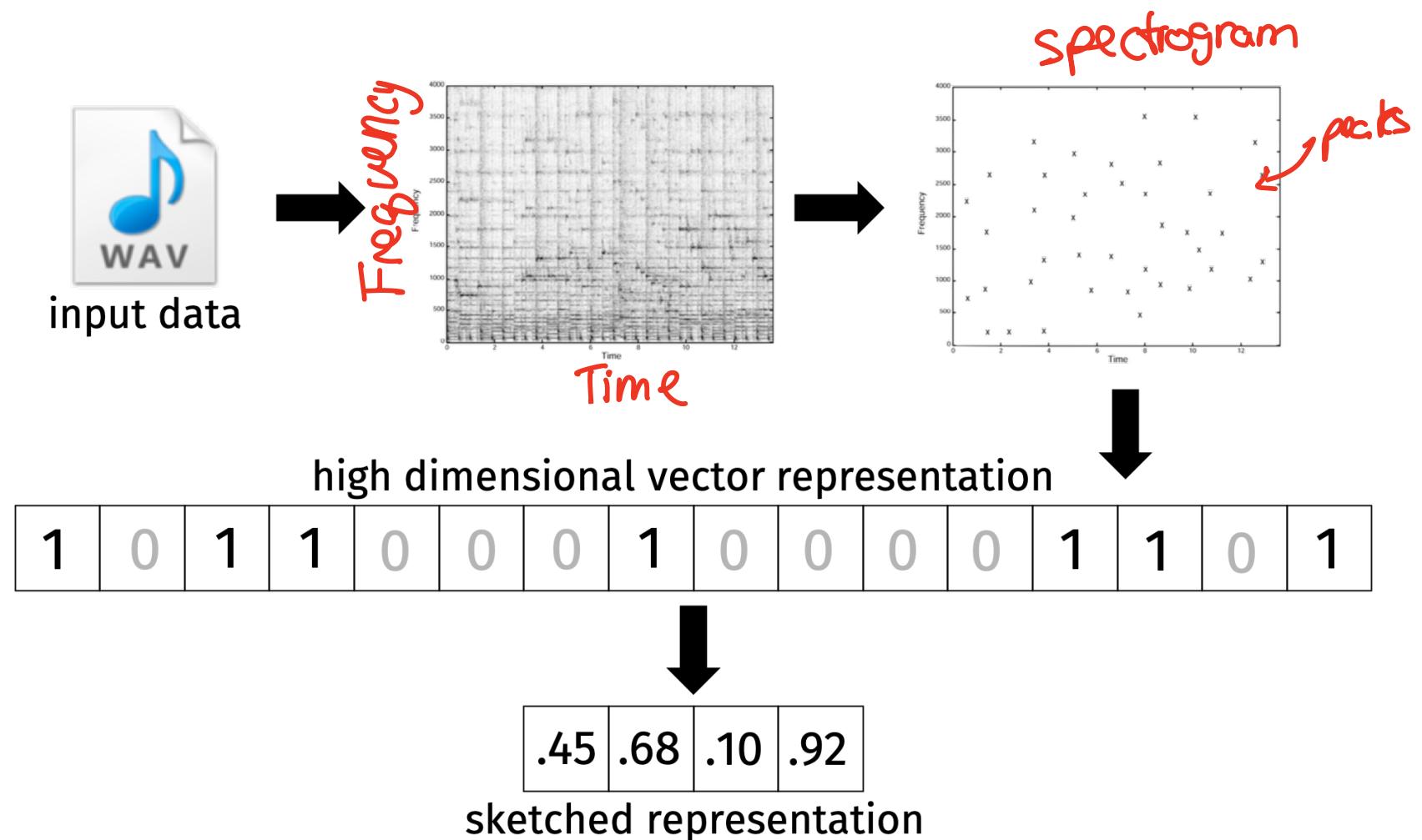
$$\|x\|_2^2 = |x|$$

estimate with $\Pi x, \Pi y$

Similarity Estimation

JL preserves distance,
how about "similarity"?

Shazam matches short,
noisy (hehe) clips against
huge database



Problem: Given query $q \in \mathbb{R}^d$, find similar song $y \in \mathbb{R}^d$

With n songs, $O(nd)$ space and $O(nd)$ naive search

"Sketch" into Lower Dimension

Want $c(q) \in \mathbb{R}^k$ for $k \ll d$

$c(x) \approx c(y)$ when $x \approx y$

How do we quantify?

Jaccard Similarity

$$J(x, y) = \frac{|X \cap Y|}{|X \cup Y|}$$

$$= \frac{\text{\#non-zero in common}}{\text{\#non-zero total}}$$

e.g. $x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$

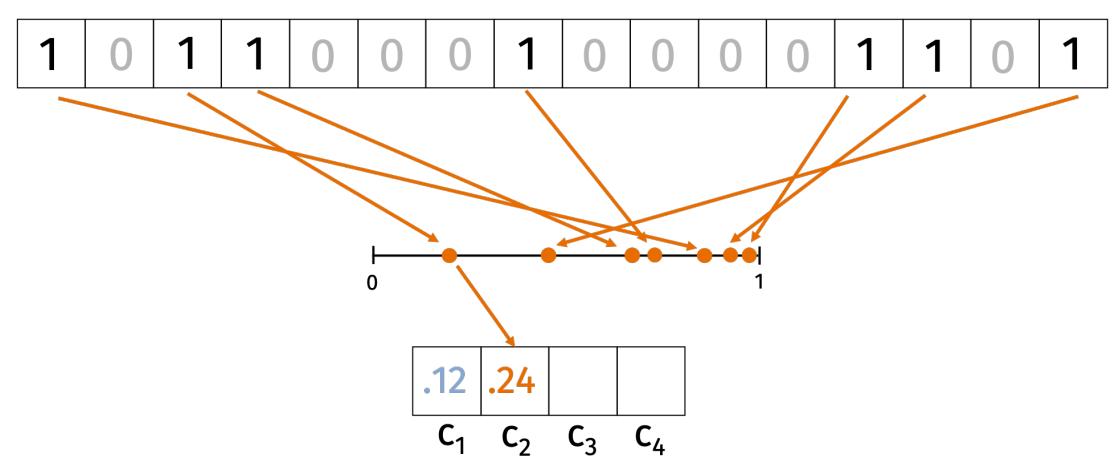
$$J(x, y) = ?$$

Also useful for

- ↳ "bag of word" documents
- ↳ cached web pages
- ↳ earthquake detection

Mish Hash

$$c: \{0,1\}^d \rightarrow \mathbb{R}^k$$



Estimate $J(x,y)$ using c_i

$$\begin{aligned}\hat{J}(x,y) &= \frac{1}{k} \sum_{i=1}^k \mathbb{I}[c_i(x) = c_i(y)] \\ &= \frac{1}{k} \langle c(x), c(y) \rangle\end{aligned}$$

$$\mathbb{E}[\hat{J}(x,y)] = ?$$

$$c_i = \min_{j \in \{1, \dots, d\}: x_j = 1} h_i(j)$$

$$\text{Var}(\hat{J}(x,y)) \leq ?$$

$$\Pr(c_i(x) = c_i(y)) = ?$$

Chebyshov's

$$\Pr(|\hat{J}(x, y) - \mathbb{E}[J(x, y)]| \geq \alpha \sqrt{\text{Var}(\tilde{J}(x, y))}) \leq \frac{1}{\alpha^2}$$

$$\epsilon = \alpha \sqrt{\text{Var}(\tilde{J}(x, y))} \quad \frac{1}{\alpha^2} \stackrel{\text{want}}{=} S$$

$$k = ?$$

$$J(x, y) - \epsilon \leq \hat{J}(x, y) \leq \hat{J}(x, y) + \epsilon \text{ wp } 1/\delta$$

using biased coin theorem

- ↳ heads if $c_i(x) = c_i(y)$
- ↳ bias $b = J(x, y)$

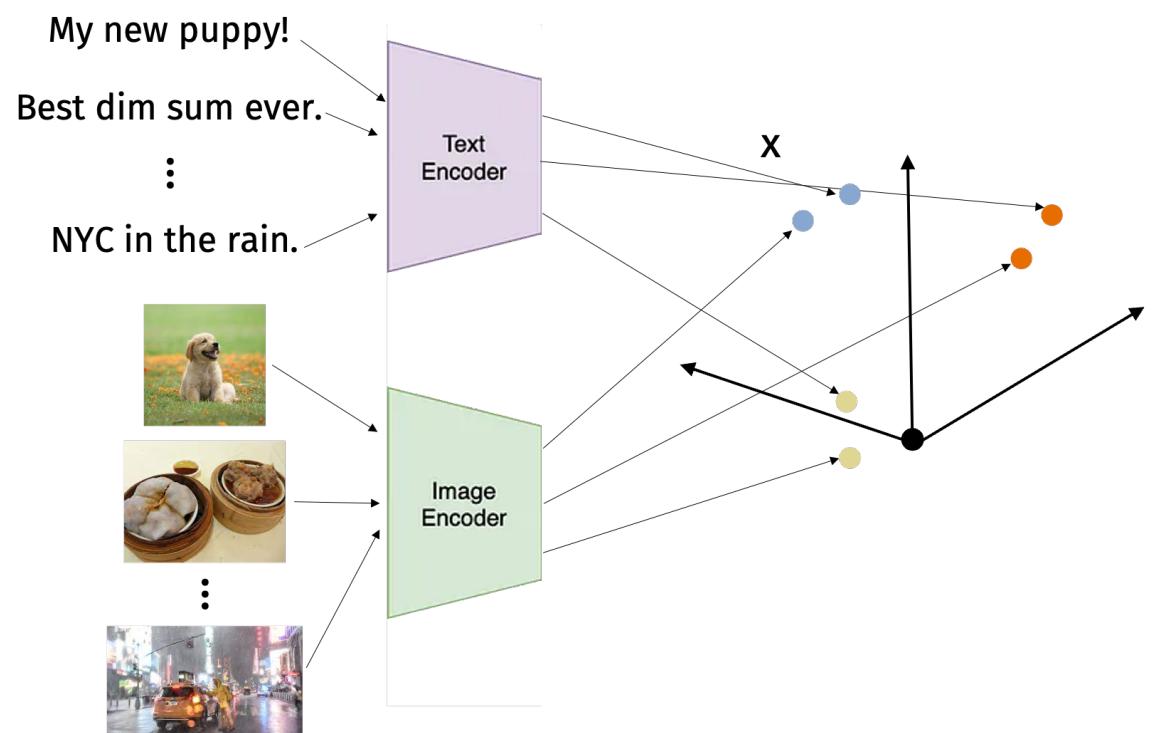
$$\text{Get } k = O\left(\frac{\log 1/\delta}{\epsilon^2}\right)$$

$O(d) \rightarrow O(k)$ compute (approx) similarity

$O(dn) \rightarrow O(kn)$ naive search

How do we find similar points faster?

useful for CLIP



Locality sensitive Hashing

- ↪ Hash function $h: \{0,1\}^d \rightarrow \{1, \dots, m\}$
- ↪ Similarity function s e.g. Jaccard
- ↪ h is locally sensitive if

$$\Pr(h(x)=h(y)) = \begin{cases} \text{large} & \text{when } x \approx y \\ \text{small} & \text{when } x \not\approx y \end{cases}$$

Our approach:

$$c: \{0,1\}^d \rightarrow [0,1] \quad \text{single MinHash}$$

$$g: \mathbb{R} \rightarrow \{1, \dots, m\} \quad \text{uniform hash function}$$

$$h(x) = g(c(x))$$

$$h(x) = h(y) \text{ when}$$

$$(1) c(x) = c(y) \text{ or}$$

$$(2) c(x), c(y) \text{ happen to hash to same cell}$$

$$\Pr(h(x)=h(y))$$

$$= \Pr(c(x)=c(y)) \cdot 1$$

$$+ \Pr(c(x) \neq c(y)) \cdot \frac{1}{m}$$

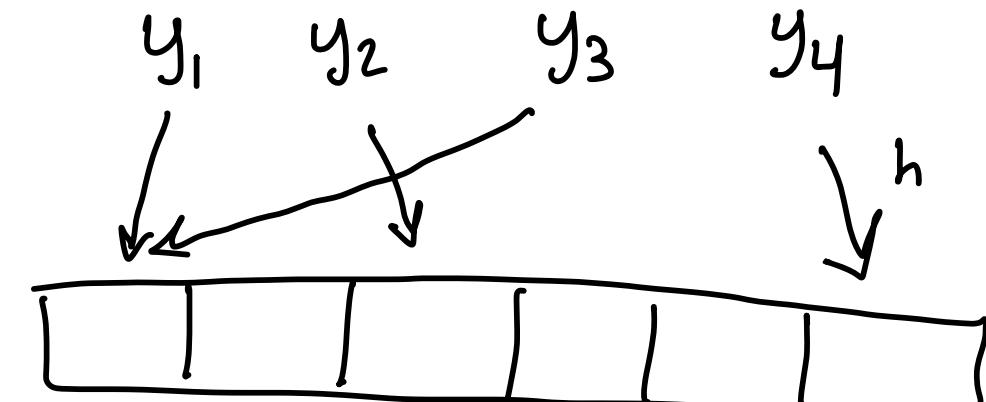
$$\leq J(x,y) + \frac{1}{m}$$

Preprocessing

Choose h in terms of g, c

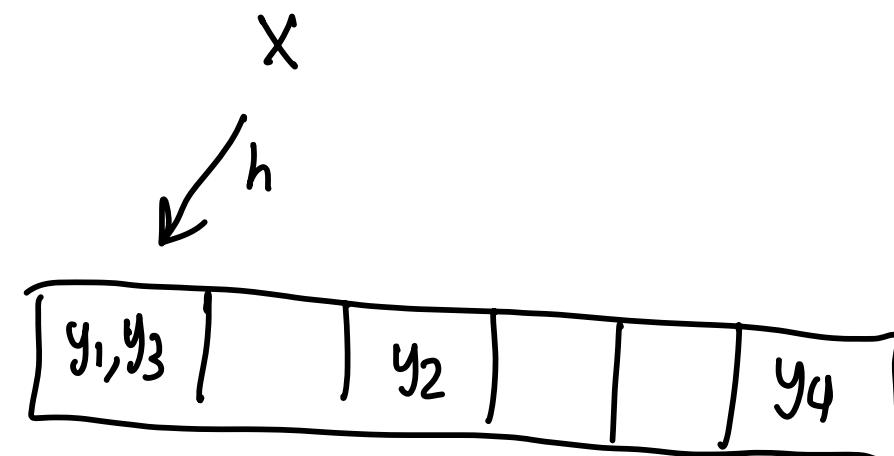
Create a table with m slots

For each vector (song), we compute $h(y)$ and store in corresponding slot



Query

Compute $h(x)$ and look in corresponding cell for similar songs



Repeat with t tables

Two questions:

- ↪ False negative: What's the probability we don't find a close vector?
- ↪ False positive: What's the probability we find a far vector?

$$\begin{aligned}\Pr(\text{find } y) &= 1 - \Pr(y \text{ not in table})^t = 1 - \Pr(h_i(x) \neq h_i(y))^t \\ &= 1 - (1 - J(x, y))^t\end{aligned}$$

when $J(x, y) = .4$ and $t=10$, $\Pr(\text{find } y) = 1 - (1 - .4)^{10} \approx .99$ ☺

when $J(x, y) = .2$ and $t=10$, $\Pr(\text{find } y) = 1 - (1 - .2)^{10} \approx .89$ ☹

Our approach:

$$c_1, \dots, c_r : \{0,1\}^d \rightarrow [0,1] \quad \text{single MinHash}$$

$$g : [0,1]^r \rightarrow \{1, \dots, m\} \quad \text{uniform hash function}$$

$$h(x) = g(c_1(x), c_2(x), \dots, c_r(x))$$

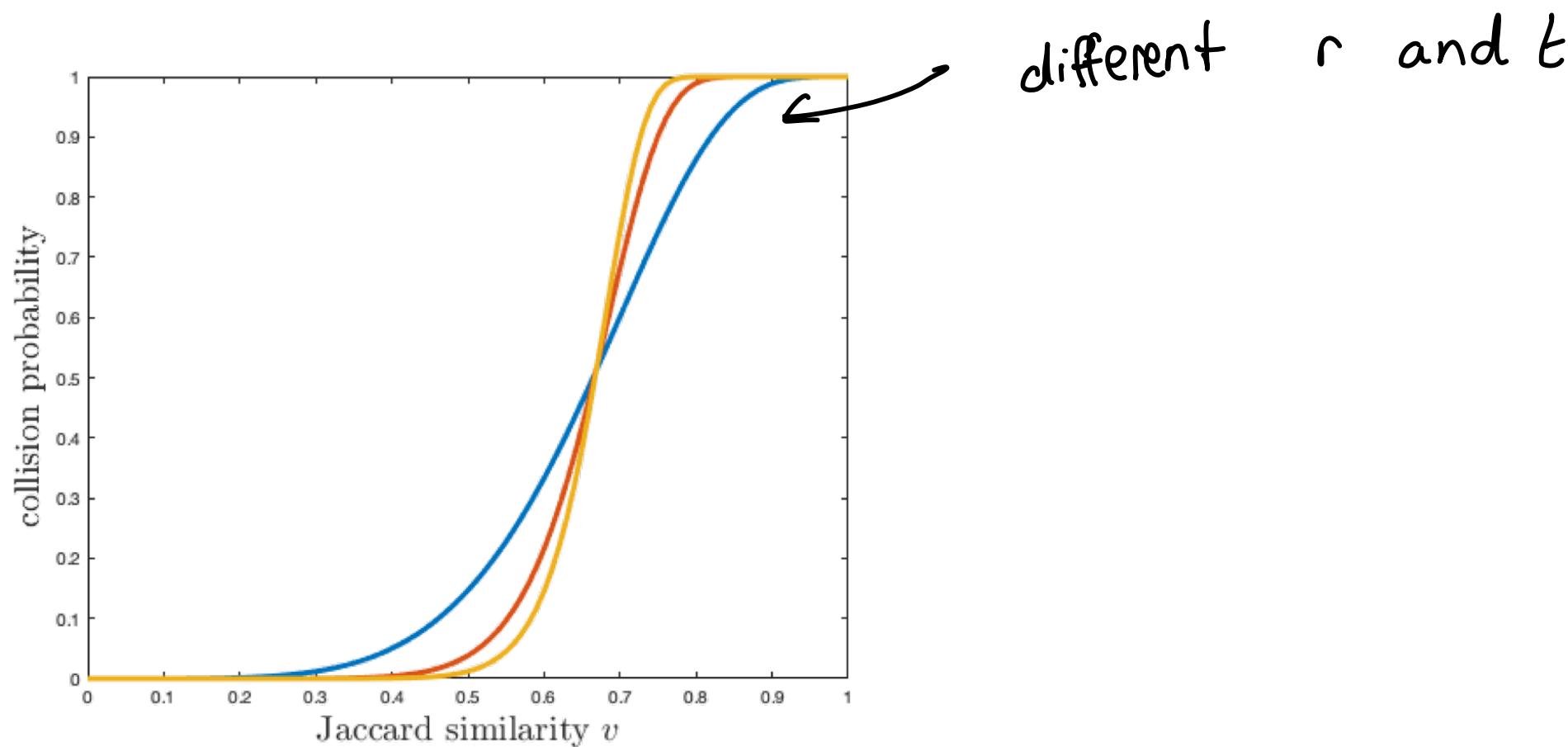
$$\begin{aligned} \Pr(h(x) = h(y)) &\leq \Pr(c_1(x) = c_1(y), \dots, c_r(x) = c_r(y)) + 1 \cdot \frac{1}{m} \\ &\stackrel{\text{indep}}{=} \Pr(c_i(x) = c_i(y))^r + \frac{1}{m} \\ &= J(x,y)^r + \frac{1}{m} \end{aligned}$$

$$\begin{aligned} \Pr(\text{find } y) &= 1 - \Pr(y \text{ not in table})^t = 1 - \Pr(h_i(x) \neq h_i(y))^t \\ &= 1 - (1 - J(x,y))^t \end{aligned}$$

$$\Pr(\text{find } y) = 1 - (1 - J(x,y))^r t$$

when $J(x,y) = .4$ and $t=10$, $r=2$, $\Pr(\text{find } y) = 1 - (1 - .4^2)^{10} \approx .83$

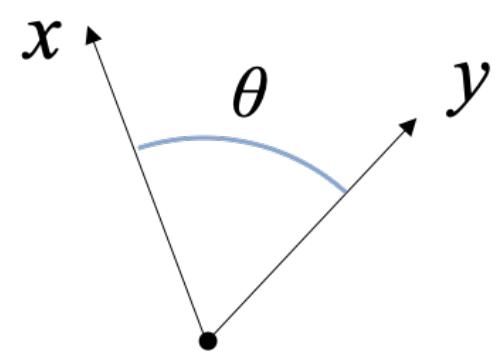
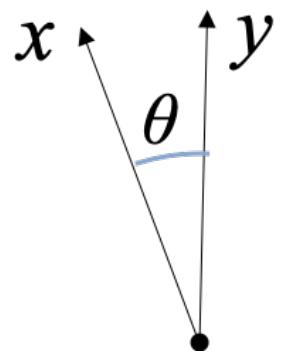
when $J(x,y) = .2$ and $t=10$, $r=2$, $\Pr(\text{find } y) = 1 - (1 - .2^2)^{10} \approx .33$



We can do the same thing for other similarity measures!

Cosine similarity

$$\cos(\theta(x, y)) = \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}$$



"inverse to distance"

$$\begin{aligned} \|x - y\|_2^2 &= \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y \rangle \\ &= \|x\|_2^2 + \|y\|_2^2 - 2 \cos \theta(x, y) \|x\|_2 \|y\|_2 \end{aligned}$$

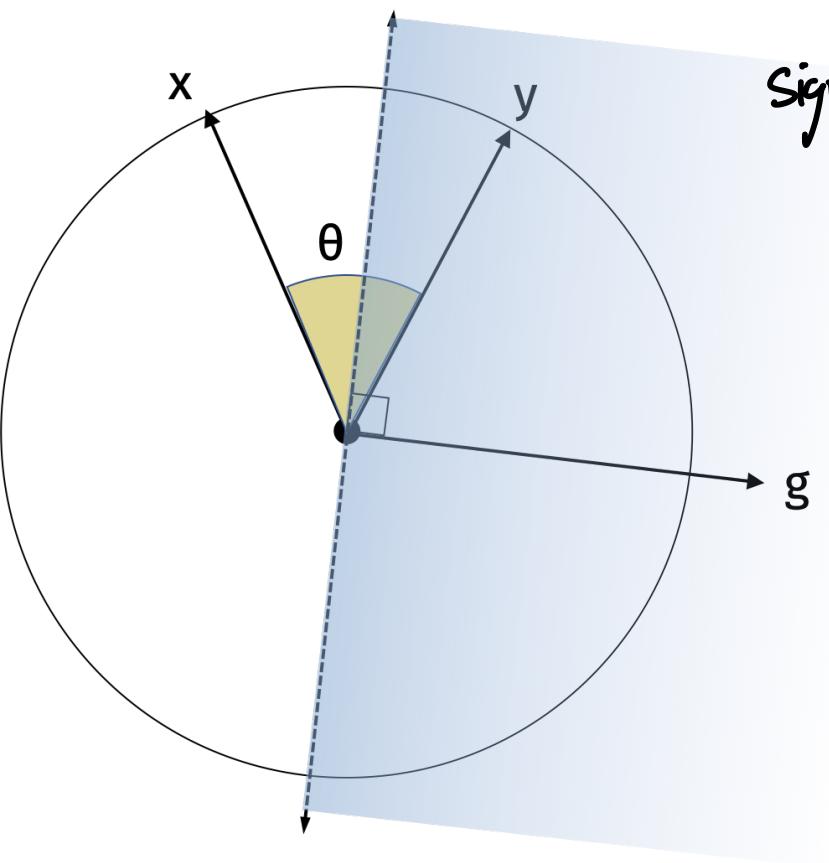
Let $g_1, \dots, g_r \in \mathbb{R}^d$ be random vectors with entries $\sim \mathcal{N}(0, 1)$

$$f: \mathbb{Z}^{-1, 1} \rightarrow \{1, \dots, m\} \text{ hash}$$

$$h: \mathbb{R}^q \rightarrow \{1, \dots, m\}$$

$$h(x) = f(\text{sign}(\langle g_1, x \rangle), \dots, \text{sign}(\langle g_r, x \rangle))$$

$$\Pr(\text{sign}(\langle g_i, x \rangle) = \text{sign}(\langle g_i, y \rangle)) = ?$$



$\text{sign}(\langle g, x \rangle)$ is which side of hyperplane

$$\Pr(\text{different sides}) =$$

$$\Pr(\text{hyperplane between}) = \frac{2\theta}{2\pi} = \frac{\theta}{\pi}$$

$$\Pr(\text{same}) = 1 - \frac{\theta}{\pi}$$

$$\Pr(\text{sign}(g_i, x) = \text{sign}(\langle g_i, y \rangle)) = 1 - \frac{\theta}{\pi}$$

$$\begin{aligned} \Pr(\text{find } y) &= 1 - \Pr(y \text{ not in table})^t = 1 - \Pr(h_i(x) = h_i(y))^t \\ &= 1 - \left(1 - \left(1 - \frac{\theta}{\pi}\right)^r\right)^t \end{aligned}$$