L7: Approximate Nearest Neighbors

Jeff M. Phillips

January 30, 2019

Edit Distance

defined between 2 short strings

operations (delete, add, replace) to turn one string into another

dedit (mines , smiles)

+ smines
2 smies
3 smiles

works only if dist small

No LSH

data $a, b \in \mathbb{R}^d$

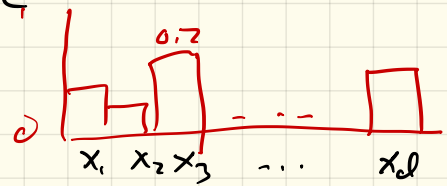
$\rightarrow L_p$

$\rightarrow \bar{a} \leftarrow \frac{a}{\|a\|_2}$ Cosine, Angular

$\rightarrow \tilde{a} \leftarrow \frac{a}{\|a\|_1}$ ensure each coord $a_i \geq 0$

$$\Delta^{d-1} = \{ x \in \mathbb{R}^d \mid \|x\|_1 = 1, \text{ each } x_i \geq 0 \}$$

$$\Delta_0^{d-1} = \{ x \in \mathbb{R}^d \mid \|x\|_1 = 1, \text{ each } x_i > 0 \}$$



$\mathbb{P}_r[x = 3] = 0.2$

$\sum_{i=1}^{d} x_i = 1$

probability dist.

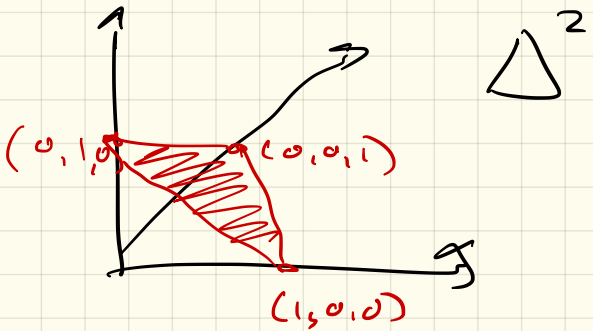
KL Divergence

$q, p \in \Delta_0^{d-1}$ ← no $p_i = 0$

$$d_{KL}(p, q) = d_{KL2}(p \| q)$$

$$= \sum_{i=1}^{d} p_i \ln\left(\frac{p_i}{q_i}\right)$$

Jensen-Shannon



Word Vector Embeddings

large corpus text (entire web)

"The story said, ..."

$v_{story} \in \mathbb{R}^{300}$

Use Euclidean dist.

Use Cosine Sim

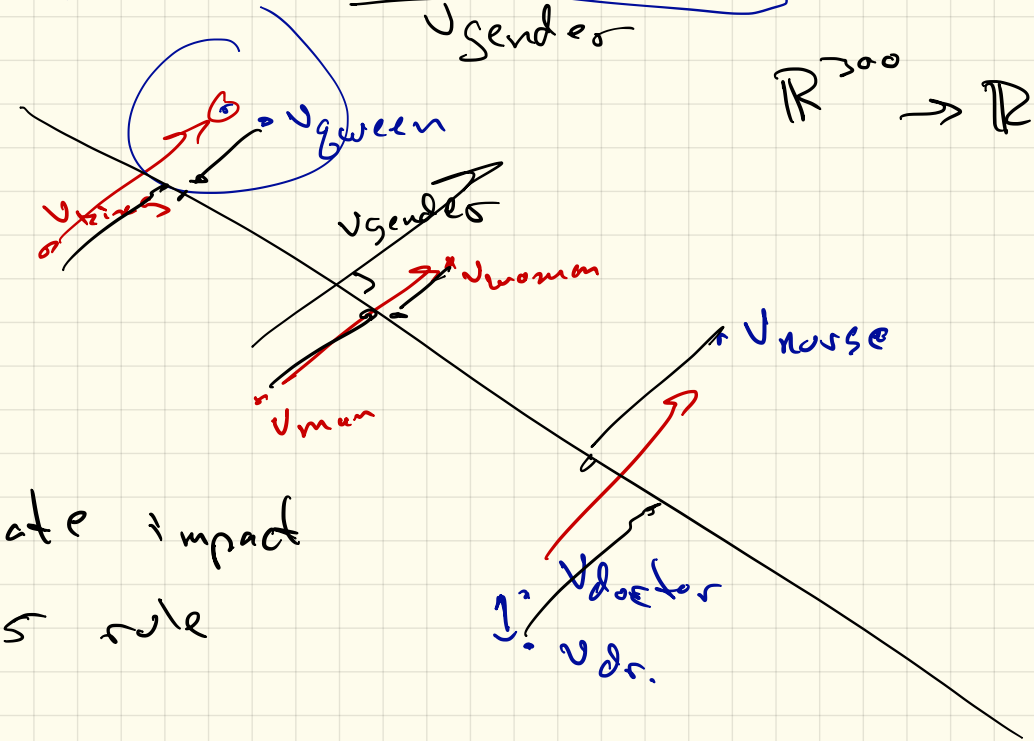
text → vectors (dim = 1,000,000 word count in neighborhood) → define PPMI → maps to $\mathbb{R}^{300}$ → v_i

$\max\left(0, \frac{P(i,j)}{P(i)P(j)}\right)$

$v_{king} + (v_{woman} - v_{man}) \approx v_{queen}$

v_{gender}

$\mathbb{R}^{300} \to \mathbb{R}^{299}$

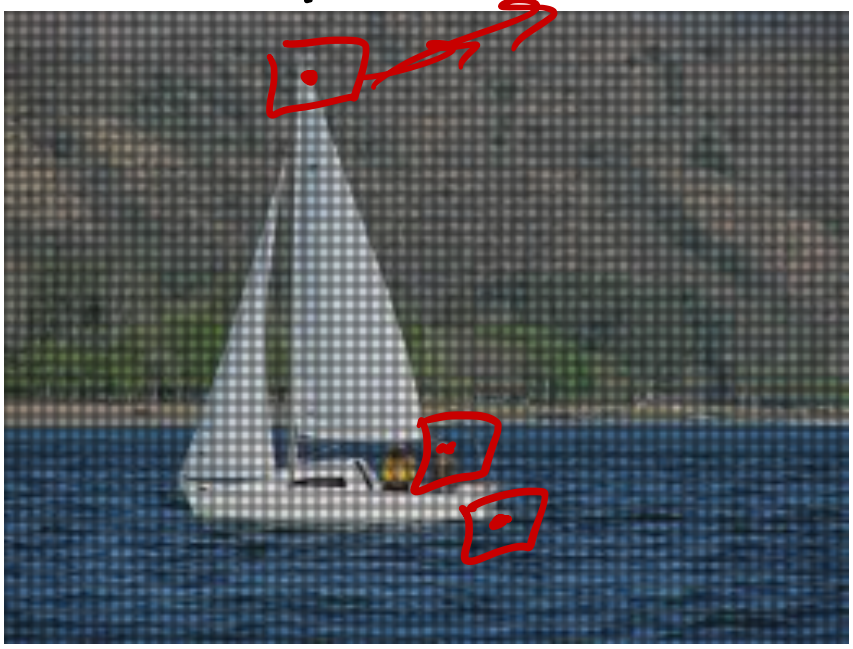


disparate impact

3/5 rule

Images and SIFT Features

Generate in $\mathbb{R}^d$

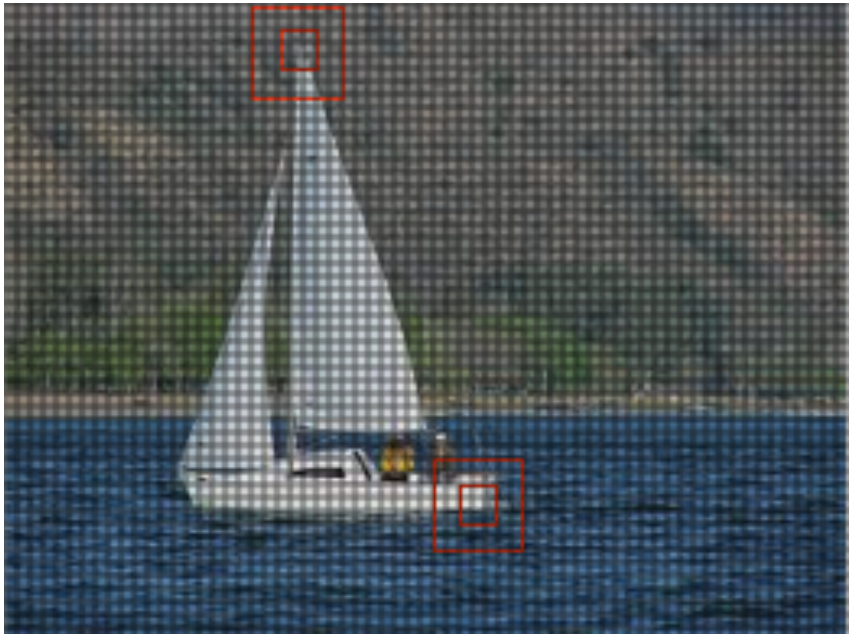


N1	N2	N3
N8	X	N4
N7	N6	N5

$\rightarrow \mathbb{R}^{200 \cdot 300 \cdot 3}$ 20 yrs ago

$\rightarrow$ SIFT vector $\rightarrow \mathbb{R}^{256}$ Euclidean 25 yrs ago

Images and SIFT Features



N1	N2	N3
N8	X	N4
N7	N6	N5

Quickly Find Nearest Neighbors

Large $P \subset \mathbb{R}^d$ $|P| = n$ ← large

→ Given query $q \in \mathbb{R}^d$

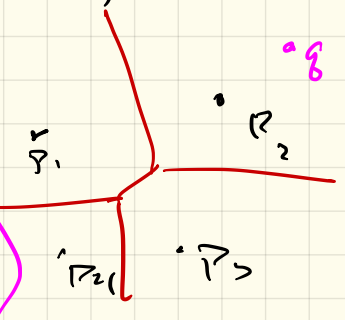
→ quickly find

$$\phi_P(q) = \arg\min_{p \in P} \|pq\|$$

- LSH binary search on T
- $d=1$ binary tree $O(\log n)$
- $d=2$ Voronoi Diagram

$O(\log n)$

size VD $O(n^{\lceil d/2 \rceil})$



$d \geq 3$ Aprox NN

Find $p \in P$ s.t.

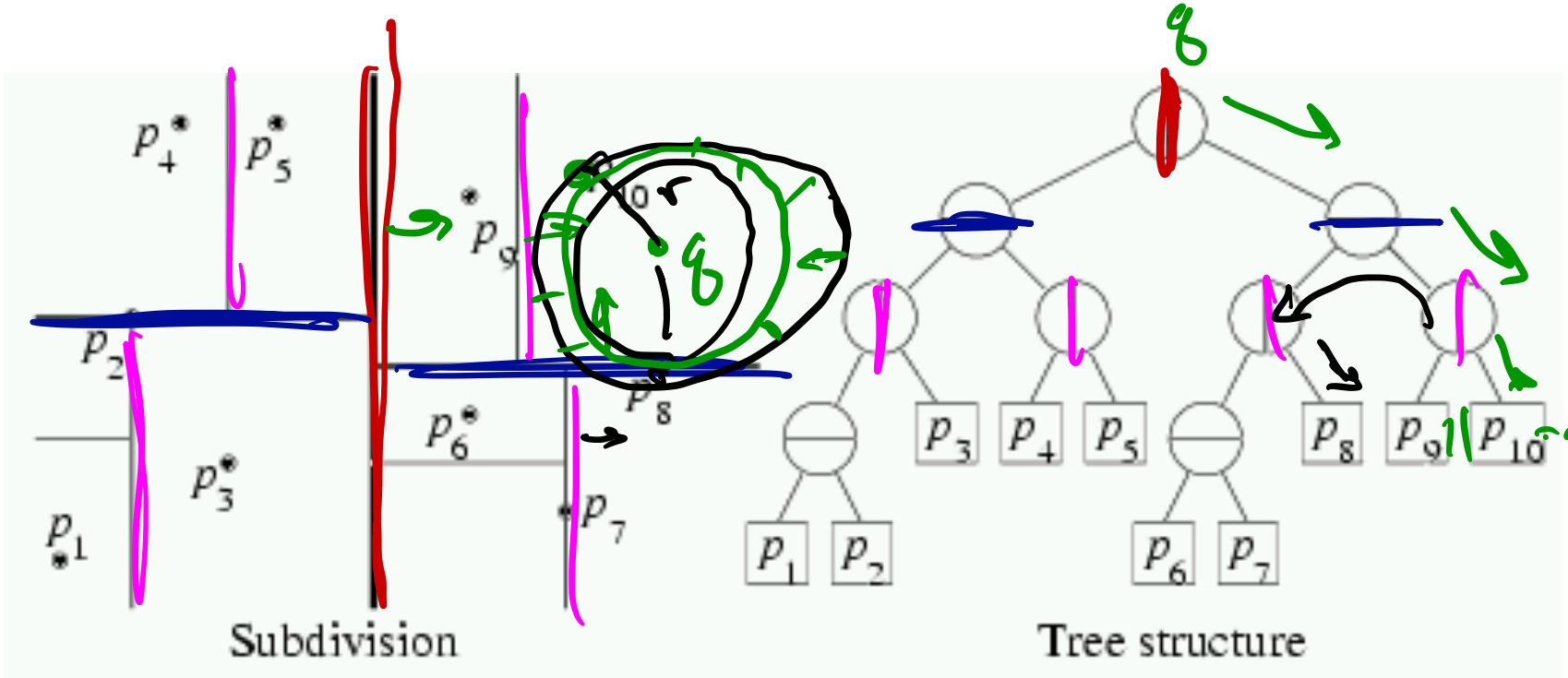
$$\|p - q\| \leq (1+\varepsilon)\|p^* - q\|$$

$$p^* = \phi_P(q)$$

$d = 3 \ldots 12$

kd-tree

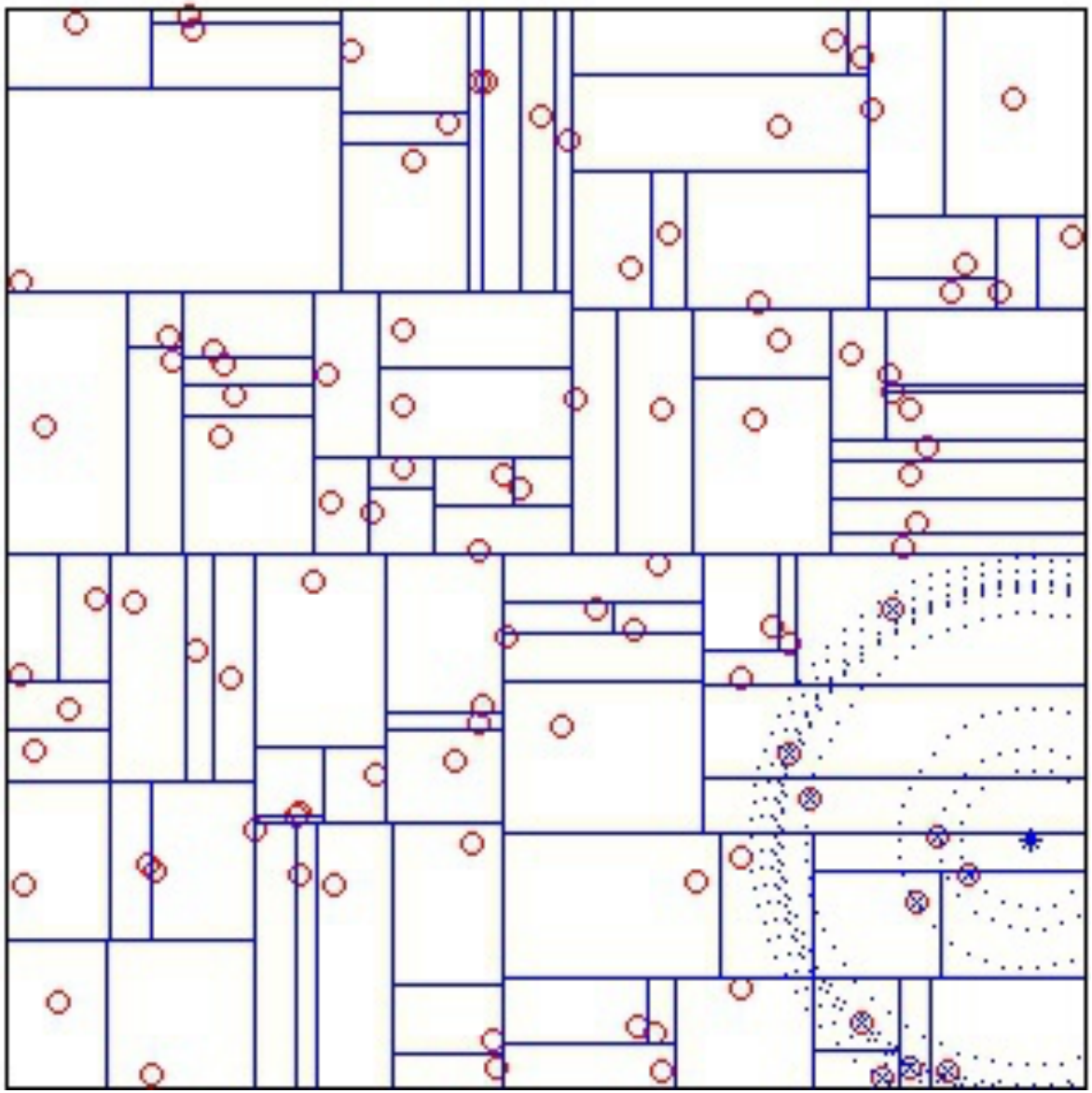
kD-Tree



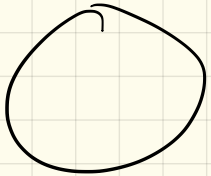
dim 2 ... 12, 20

Data-adaptive kd-tree → d=200

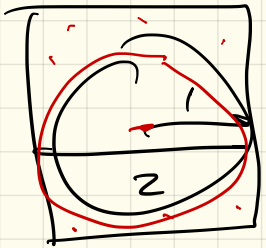
Approximate Queries on kD-Tree



vs.



ball rad 1 vol



$$\frac{\pi^{d/2}}{\Gamma(d/2+1)} \simeq \frac{\pi^{d/2}}{(d/2)!} \to 0 \text{ as } d \to \infty$$

$1 \cdot 2 \cdot 3 \cdots d/2$

box side length 2 vol

$$2^d$$