Approximate Nearest Neighbors (+SIFT)

Note Title 2/3/2016

Large Dataset P $|P| = n = 1$ million

(Q1) Which pairs $p, p' \in P$ are close

(Q2) Given query q find closest point $p \in P$ to q.

$$p^* = \underset{p \in P}{\operatorname{argmin}}\, d(p, q) = \phi_P(q)$$

High Dimensional Euclidean Data

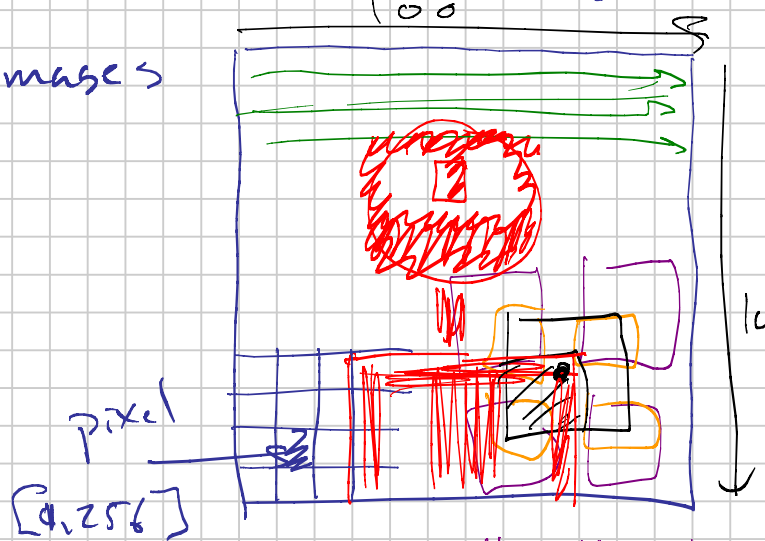
- time series: price stock temperature

  1 2 3
  15 15 16 18 ... 35
  10

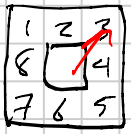
- machine learning kernel $K(x, y)$ → RKHS
  ↳ lrfd

- Images

  $\mathbb{R}^m$ $m = \frac{1}{\varepsilon^2}$



  - Vector $d =$ $100 \times 100 \to 10{,}000$ (if pre-aligned)
  - SIFT (David Lowe & coord)



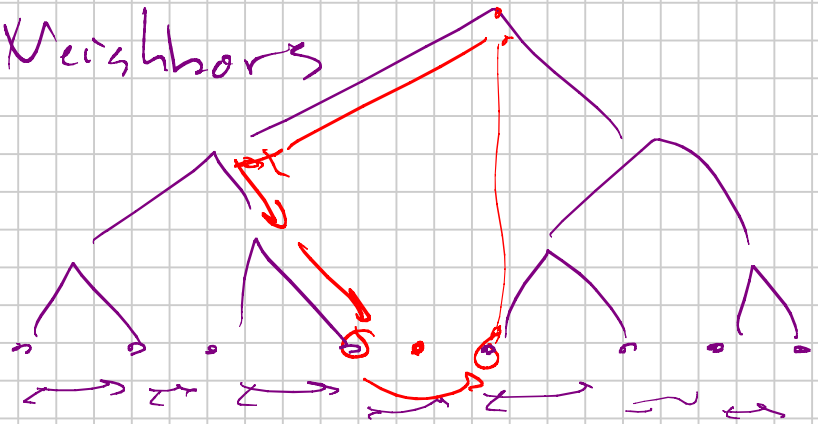
4 × 4 = 16 scale

8 · 16 = 128 dimensions

invariant to scale, rotation, shift

Nearest Neighbors

$P \subset \mathbb{R}^1$



$\log n$

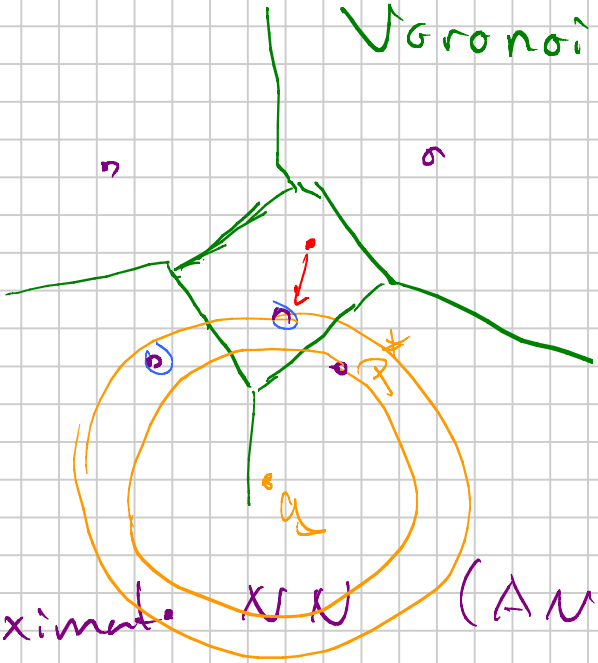
$P \subset \mathbb{R}^2$

Voronoi Diagram

size $= O(n)$

query $= O(\log n)$

construction $= O(n \log n)$



$P \subset \mathbb{R}^d$

d-dim

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

Approximate NN (ANN)

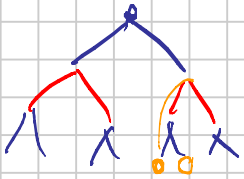
$p^* = \underset{p \in P}{\operatorname{argmin}}\, d(p, q) = \phi_P(q)$

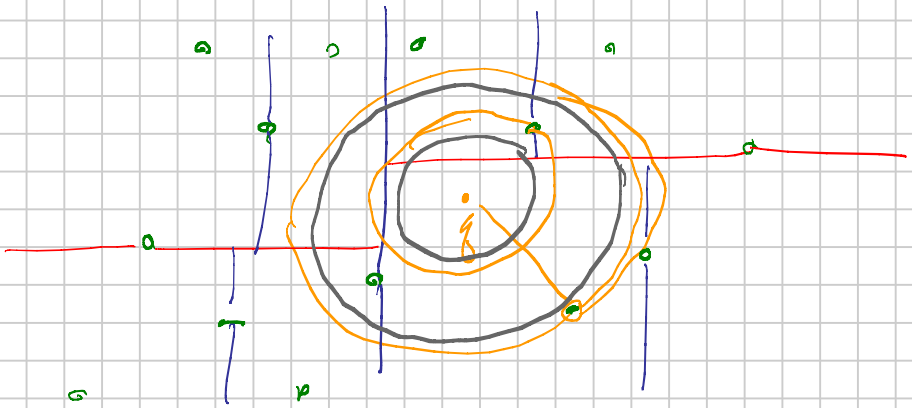
$\varepsilon \in (0, 1)$ (e.g. 0.10)

Goal: Find $\hat{p}$ s.t. $d(\hat{p}, q) \leq (1+\varepsilon)\, d(p^*, q)$

Medium Dimensions $d \in [3-12]$

k-d tree

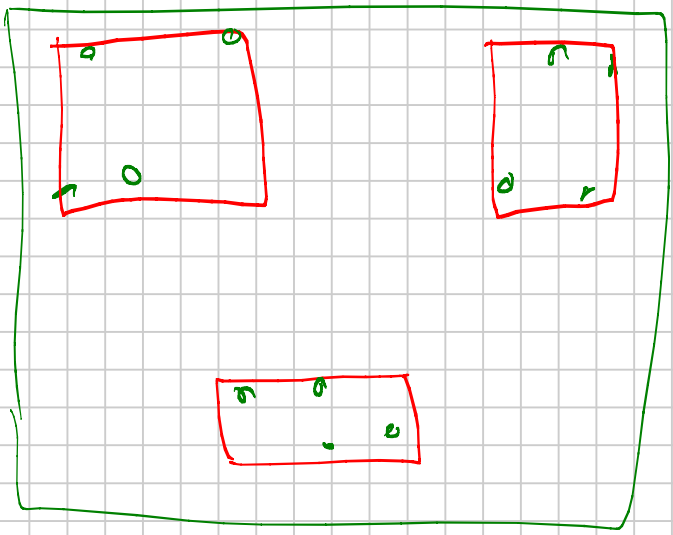




ANN

- R-tree

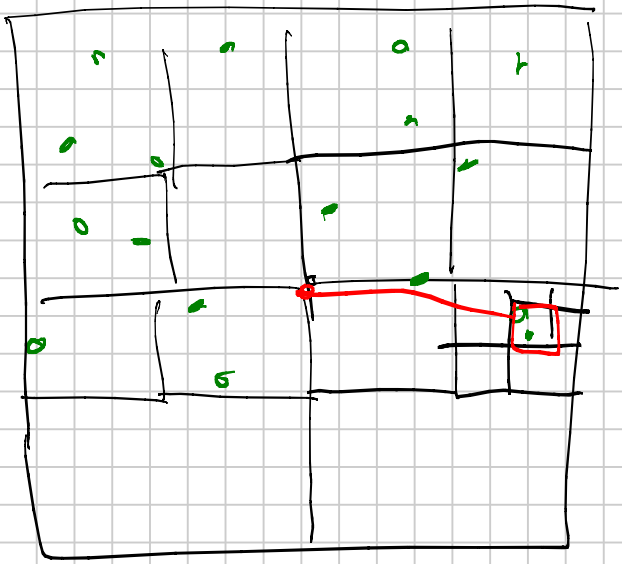
  R = rectangle



- Quad-tree (compressed)

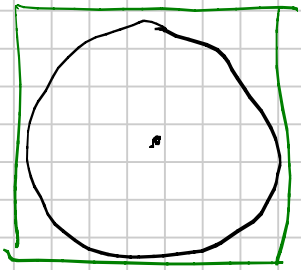
  $O(n \log n + \frac{1}{\varepsilon^d})$

  $O((\log n + \frac{1}{\varepsilon})^d)$



High-Dimensions $d > 20$

"curse of dimensionality"



$B(d, \text{rad}=1)$ ball in $\mathbb{R}^d$

$$\text{Vol}(B(d,1)) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \approx \frac{\pi^{d/2}}{(d/2)!} \leq 1$$

(radius ≤ 1)

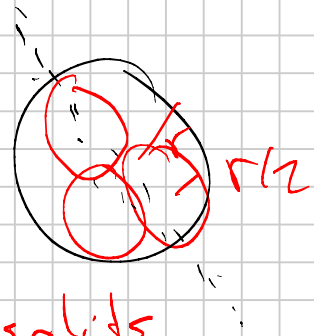
$[-1, 1]^d$ $\quad \text{Vol}(\text{cube}(d, \text{rad}=1)) = 2^d$

- ANN $d \in [2, 20?]$
- LSH $n >$ very big

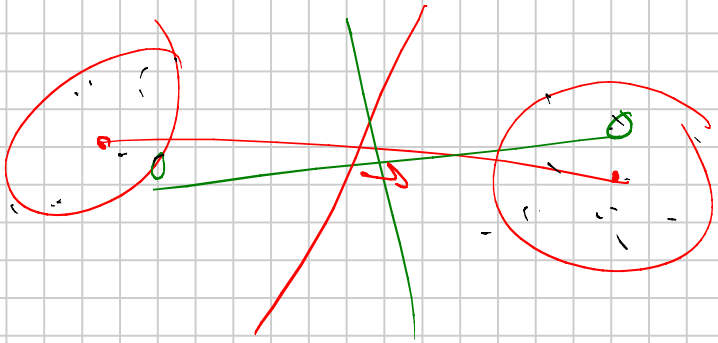
- Mods to kd-tree $d \approx 100\text{-}200$

inherently low dimensional

bounded doubling dimension



- Build kd tree but smarter splits



- cluster
  2 clusters
- Random Rotations
  do few
  pick best