

L5: Nearest Neighbor Search + HNSW

Sep 3, 2023
Data Mining



Jeff M. Phillips

Nearest Neighbor Search

Data set

distance

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

$$D: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$$

$$D(a, b) = \|a - b\|$$

$$1. \text{NN}(g, X)$$

$$x^* = S_X(g)$$

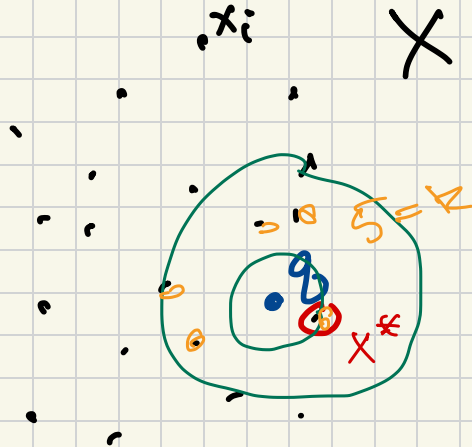
$$2. k\text{-NN}(g, X)$$

$$\{x_1, \dots, x_k\} = S_X(g, k)$$

query $g \in \mathbb{R}^d$

$$x^* = \Phi_X(g) = \underset{x \in X}{\operatorname{argmin}} D(x, g)$$

preprocess $X \rightarrow S_X$
data structure

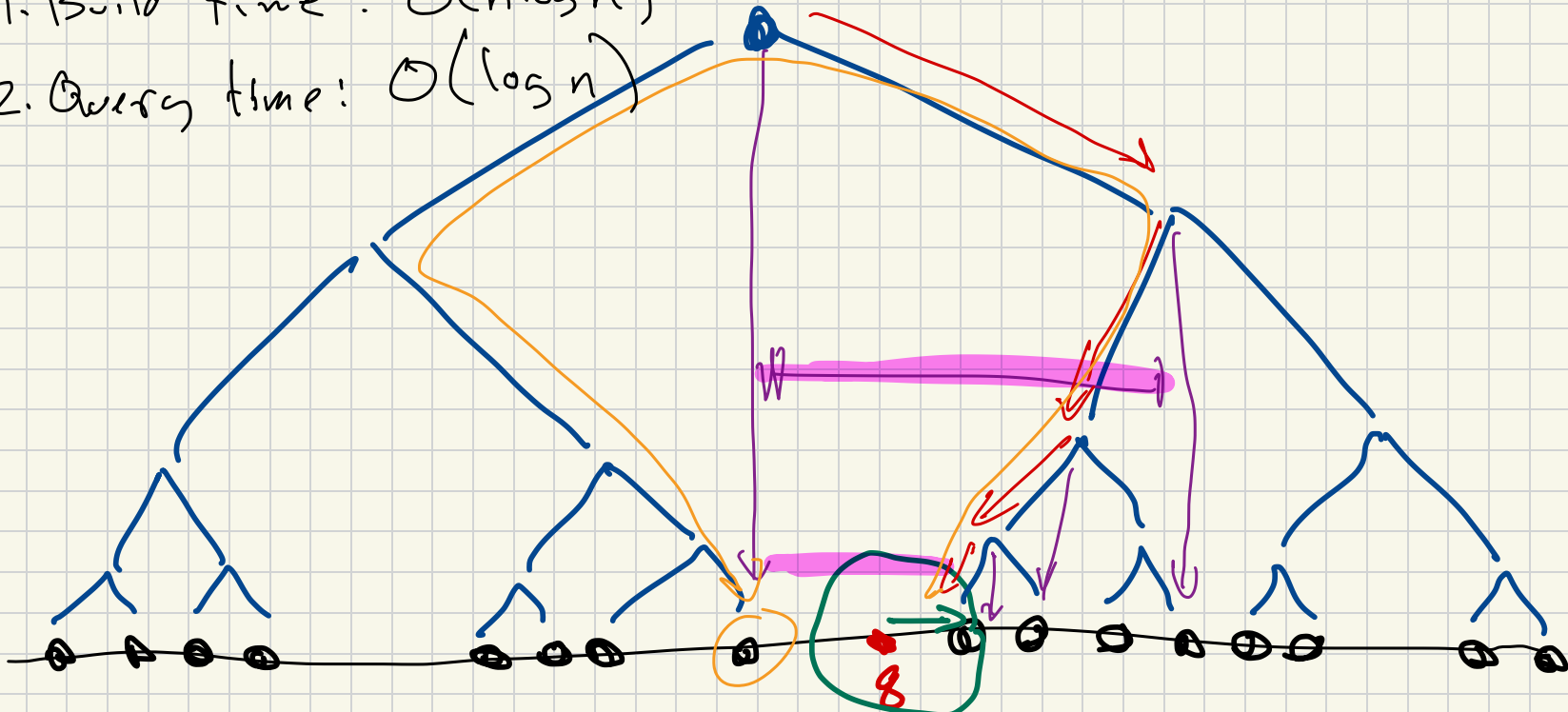


1-Dim N N Search

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

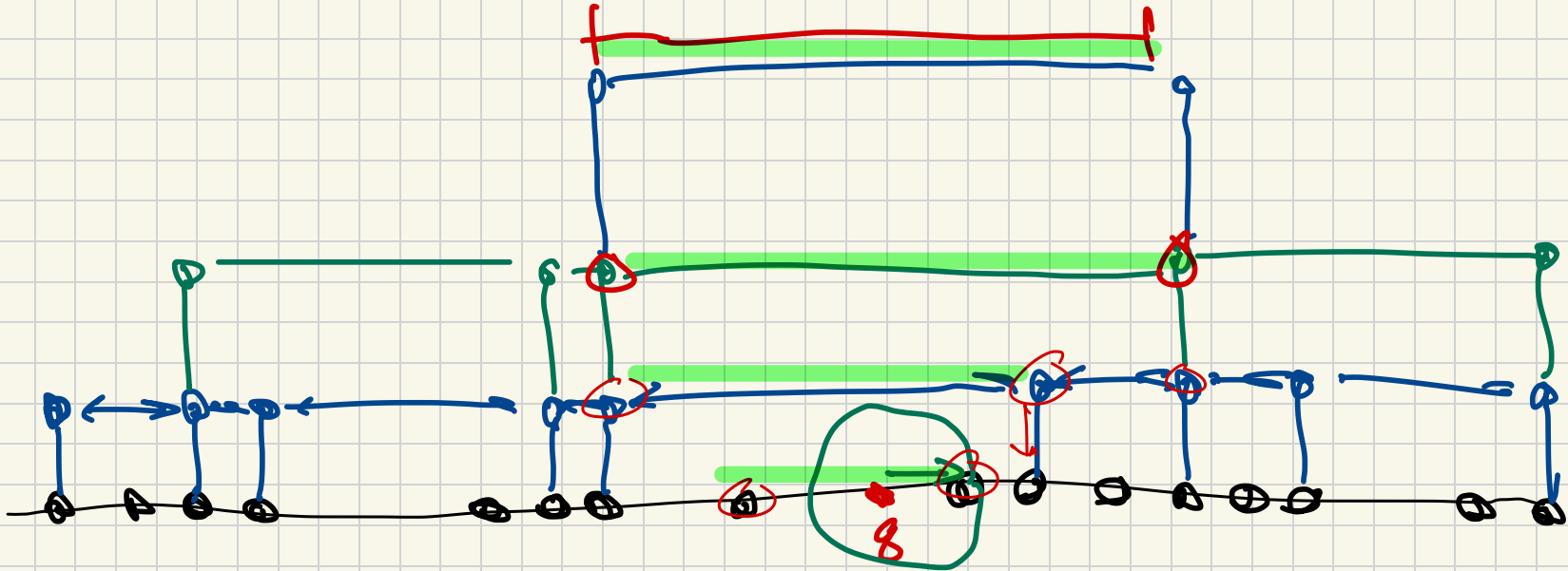
giving $g \in \mathbb{R}$

1. Build time: $O(n \log n)$
2. Query time: $O(\log n)$



Skip List

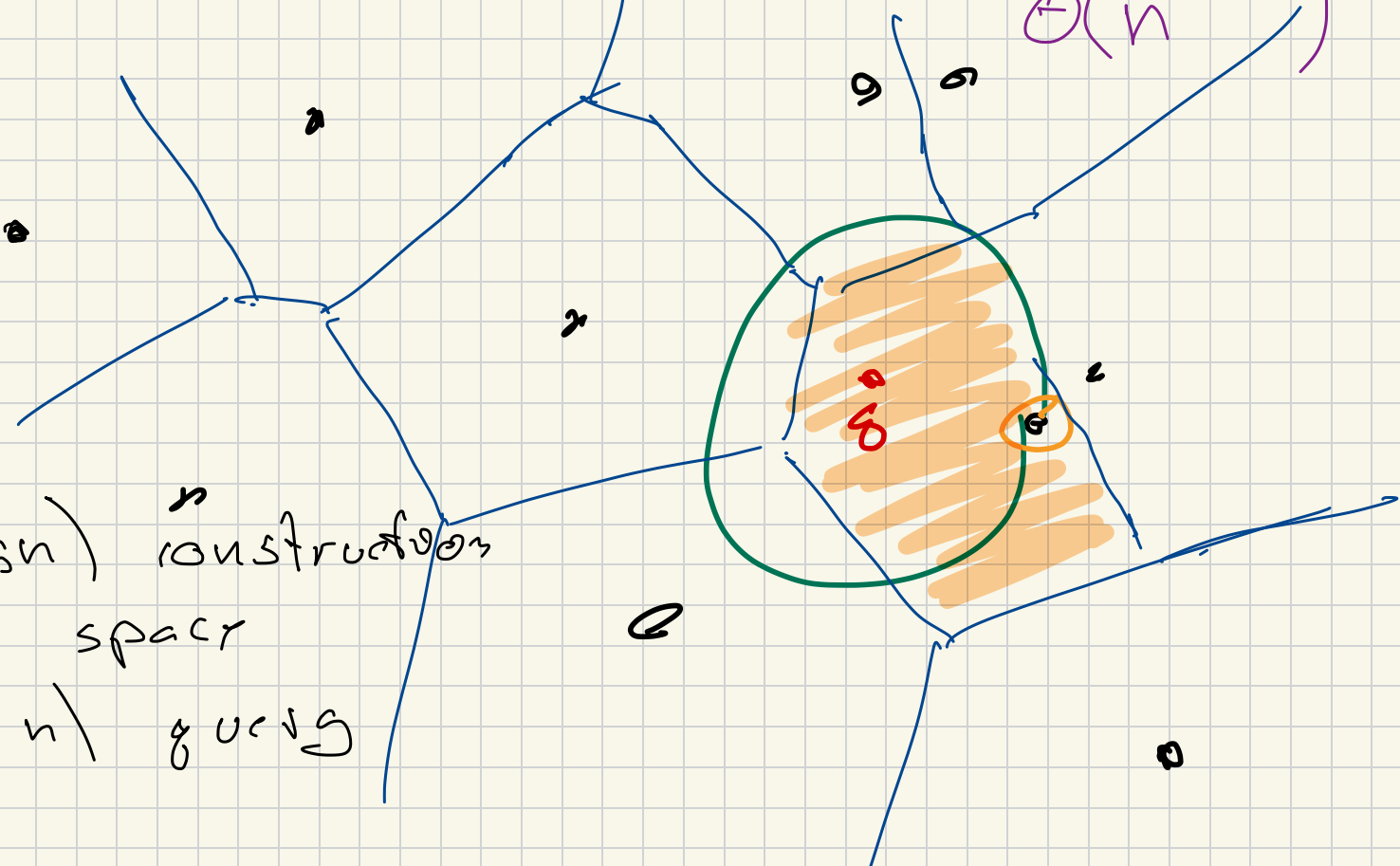
query q



2-Dim NN search

space $O(n^{d/2})$

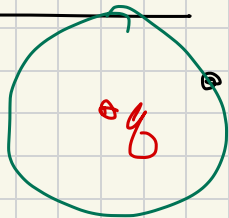
$d=2$
 $O(n \log n)$ construction
 $O(n)$ space
 $O(\log n)$ querying



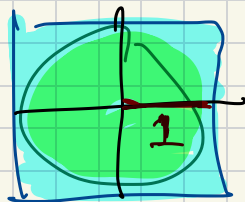
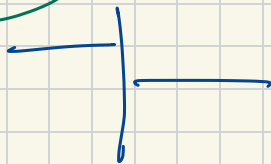
Approximate Nearest Neighbor Search

High Dimensions

Solve



Decomposition



Ball radius 1

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

$$d=2 \Rightarrow \pi$$

$$d=4 \Rightarrow \pi^2/24$$

$$d \rightarrow \infty \Rightarrow \text{Vol}(d) \Rightarrow 0$$

$$\begin{aligned} \text{Vol}(\text{Box}(\text{len}=2)) \\ = 2^d \end{aligned}$$

Locality Sensitive Hashing

Randomized Data Structure

↳ space, query time, construction time
do not depend on d .

ϵ -approx, time/space depends on ϵ .

For $\epsilon = 1$ (2-approx)

space $\tilde{O}(n^{8/7})$

query $\tilde{O}(n^{1/7})$

Theoretically
optimal.

FALCONN

Hierarchical Graph

entry point
↙ ↘

pine cone
Dist Ann
FAISS
K Graph

