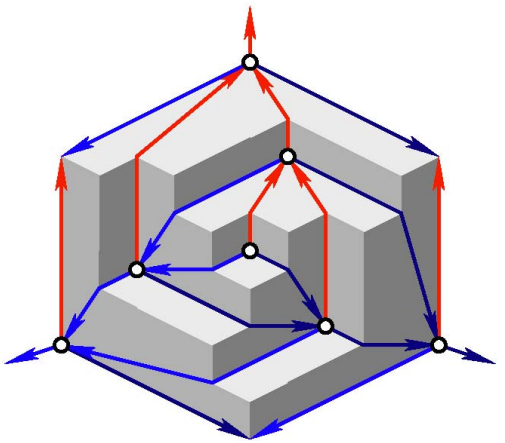


Approximate k-flat Nearest Neighbor Search

Wolfgang Mulzer
Freie Universität
Berlin, Germany

Huy L. Nguyễn
TTI
Chicago, USA

Paul Seiferth
Freie Universität
Berlin, Germany

Yannik Stein
Freie Universität
Berlin, Germany

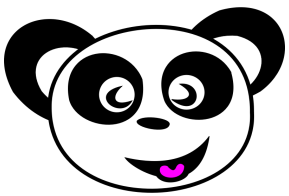
Approximate Nearest Neighbor Search (ANN)

given a set of n objects, s.t. each can be identified by d features

Approximate Nearest Neighbor Search (ANN)

given a set of n objects, s.t. each can be identified by d features

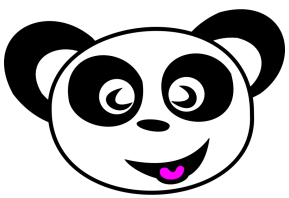
want: data structure that finds a "similiar" object for a query object

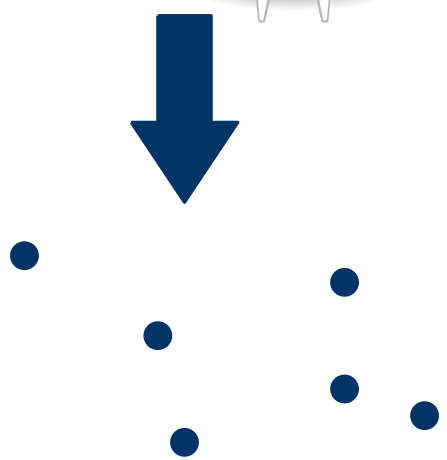


Approximate Nearest Neighbor Search (ANN)

given a set of n objects, s.t. each can be identified by d features

want: data structure that finds a "similiar" object for a query object



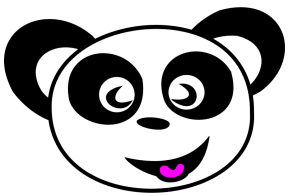


n points $P \subset \mathbb{R}^d$

Approximate Nearest Neighbor Search (ANN)

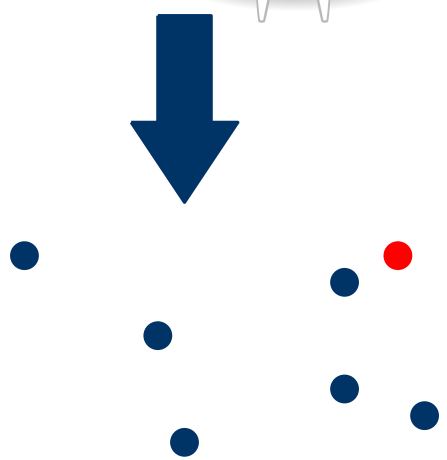
given a set of n objects, s.t. each can be identified by d features

want: data structure that finds a "similiar" object for a query object



query point q

find $p \in P$ s.t. $d(q,p) \leq cd(q,P)$, where $c > 1$ is the approx. factor

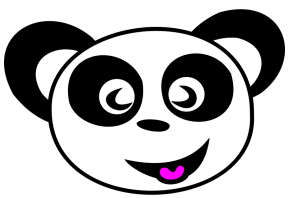


n points $P \subset \mathbb{R}^d$

Approximate Nearest Neighbor Search (ANN)

given a set of n objects, s.t. each can be identified by d features

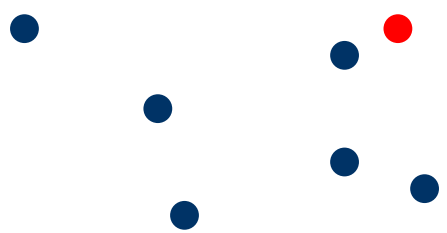
want: data structure that finds a "similiar" object for a query object



query point q

find $p \in P$ s.t. $d(q,p) \leq cd(q,P)$, where $c > 1$ is the approx. factor

plethora of applications: pattern recognition, recommendation systems dna sequencing, databases, ...

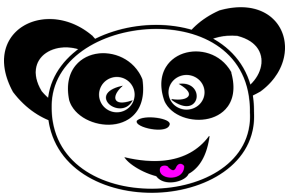


n points $P \subset \mathbb{R}^d$

Approximate Nearest Neighbor Search (ANN)

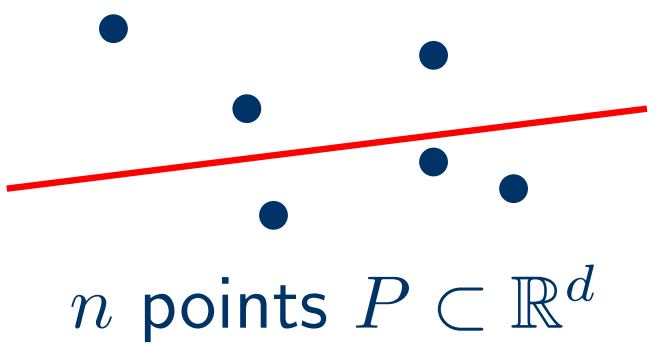
given a set of n objects, s.t. each can be identified by d features

want: data structure that finds a "similiar" object for a query object



query ~~point q~~
line ℓ

find $p \in P$ s.t. $d(q,p) \leq cd(q,P)$, where $c > 1$ is the approx. factor



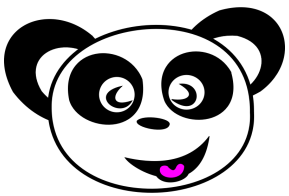
n points $P \subset \mathbb{R}^d$

plethora of applications: pattern recognition, recommendation systems dna sequencing, databases, ...

Approximate Nearest Neighbor Search (ANN)

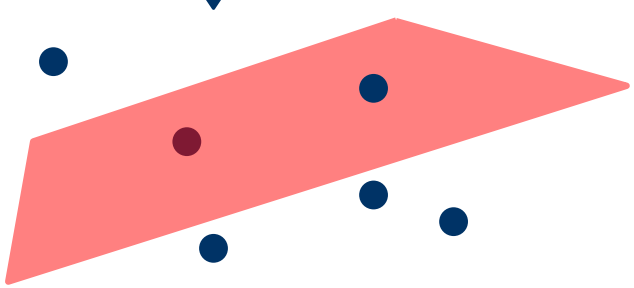
given a set of n objects, s.t. each can be identified by d features

want: data structure that finds a "similiar" object for a query object



query ~~point q~~
line ℓ k-flat f

find $p \in P$ s.t. $d(q,p) \leq cd(q,P)$,
where $c > 1$ is the approx. factor



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

plethora of applications: pattern recognition, recommendation systems dna sequencing, databases, ...

Known Results vs. Our Approach

desired approximation factor: $c = 1 + \varepsilon$, $\varepsilon > 0$

	Space	Query time	Authors
0-ANN	$d^{O(1)} n^{1+\sigma}$ $\sigma = 1/c^2$ $\sigma = \frac{7}{8}c^{-2} + O(c^{-3})$	$d^{O(1)} n^{\rho}$ $\rho = 1/c^2$ $\rho = \frac{7}{8}c^{-2} + O(c^{-3})$	Andoni, PhD Thesis Andoni, Indyk, N., Razenshteyn, SODA'14
1-ANN	$d^{O(1)} n^{O(\varepsilon^{-2}+t^{-2})}$	$d^{O(1)} n^{1/2+t}$	A., I., Krauthgamer, N., SODA'09
k-ANN	? ? ?	? ? ?	

Known Results vs. Our Approach

desired approximation factor: $c = 1 + \varepsilon$, $\varepsilon > 0$

	Space	Query time	Authors
0-ANN	$d^{O(1)} n^{1+\sigma}$ $\sigma = 1/c^2$ $\sigma = \frac{7}{8}c^{-2} + O(c^{-3})$	$d^{O(1)} n^{\rho}$ $\rho = 1/c^2$ $\rho = \frac{7}{8}c^{-2} + O(c^{-3})$	Andoni, PhD Thesis Andoni, Indyk, N., Razenshteyn, SODA'14
1-ANN	$d^{O(1)} n^{O(\varepsilon^{-2}+t^{-2})}$	$d^{O(1)} n^{1/2+t}$	A., I., Krauthgamer, N., SODA'09
k-ANN	? ? ?	? ? ?	

Our solution: decompose P into clusters

- Build cluster structure
- Build projection structure

Known Results vs. Our Approach

desired approximation factor: $c = 1 + \varepsilon$, $\varepsilon > 0$

	Space	Query time	Authors
0-ANN	$d^{O(1)} n^{1+\sigma}$ $\sigma = 1/c^2$ $\sigma = \frac{7}{8}c^{-2} + O(c^{-3})$	$d^{O(1)} n^{\rho}$ $\rho = 1/c^2$ $\rho = \frac{7}{8}c^{-2} + O(c^{-3})$	Andoni, PhD Thesis Andoni, Indyk, N., Razenshteyn, SODA'14
1-ANN	$d^{O(1)} n^{O(\varepsilon^{-2}+t^{-2})}$	$d^{O(1)} n^{1/2+t}$	A., I., Krauthgamer, N., SODA'09
k-ANN	? ? ?	? ? ?	

gen. approach

Our solution: decompose P into clusters

- Build cluster structure
- Build projection structure

Known Results vs. Our Approach

desired approximation factor: $c = 1 + \varepsilon$, $\varepsilon > 0$

	Space	Query time	Authors
0-ANN	$d^{O(1)} n^{1+\sigma}$ $\sigma = 1/c^2$ $\sigma = \frac{7}{8}c^{-2} + O(c^{-3})$	$d^{O(1)} n^{\rho}$ $\rho = 1/c^2$ $\rho = \frac{7}{8}c^{-2} + O(c^{-3})$	Andoni, PhD Thesis Andoni, Indyk, N., Razenshteyn, SODA'14
1-ANN	$d^{O(1)} n^{O(\varepsilon^{-2}+t^{-2})}$	$d^{O(1)} n^{1/2+t}$	A., I., Krauthgamer, N., SODA'09
k-ANN	? ? ?	? ? ?	

gen. approach

blackbox

Our solution: decompose P into clusters

Build cluster structure

Build projection structure

Known Results vs. Our Approach

desired approximation factor: $c = 1 + \varepsilon$, $\varepsilon > 0$

	Space	Query time	Authors
0-ANN	$d^{O(1)} n^{1+\sigma}$ $\sigma = 1/c^2$ $\sigma = \frac{7}{8}c^{-2} + O(c^{-3})$	$d^{O(1)} n^{\rho}$ $\rho = 1/c^2$ $\rho = \frac{7}{8}c^{-2} + O(c^{-3})$	Andoni, PhD Thesis Andoni, Indyk, N., Razenshteyn, SODA'14
1-ANN	$d^{O(1)} n^{O(\varepsilon^{-2}+t^{-2})}$	$d^{O(1)} n^{1/2+t}$	A., I., Krauthgamer, N., SODA'09
k-ANN	$n^{1+k\sigma/(k+1-\rho)} \log^{1/t} n$	$d^{O(1)} n^{k/(k+1-\rho)+t}$	

gen. approach

blackbox

Our solution: decompose P into clusters

Build cluster structure

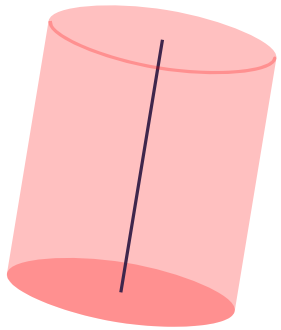
Build projection structure

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.

General Framework (Preprocessing)

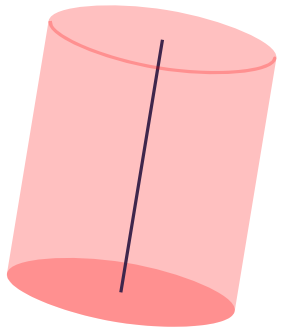
Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



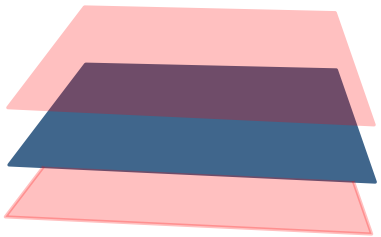
$k = 1$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



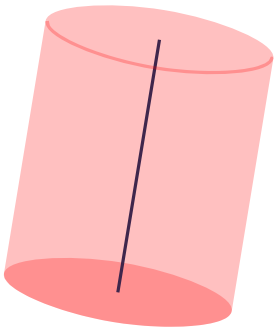
$k = 1$



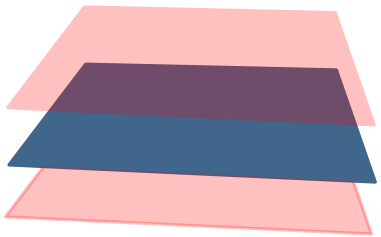
$k = 2$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.

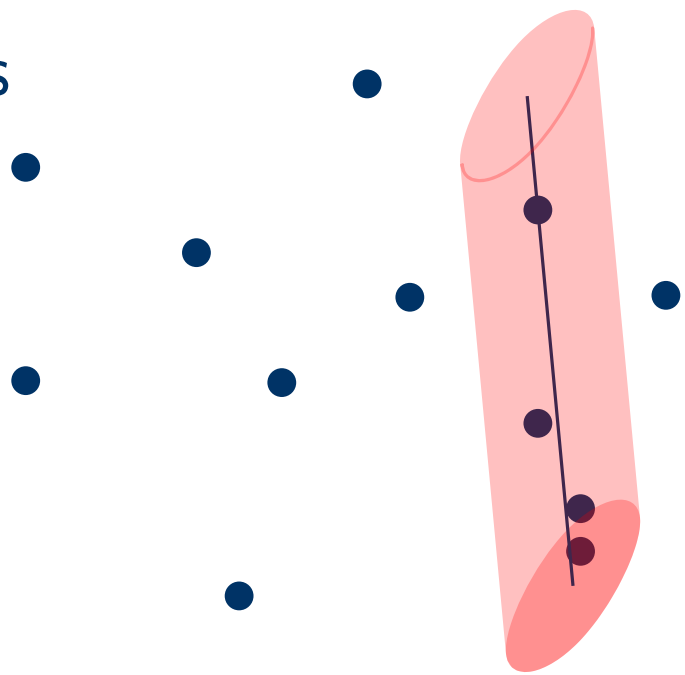


$k = 1$



$k = 2$

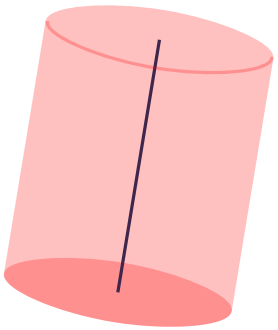
Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points



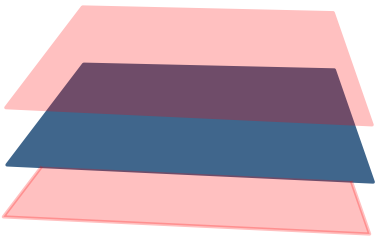
C_1

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.

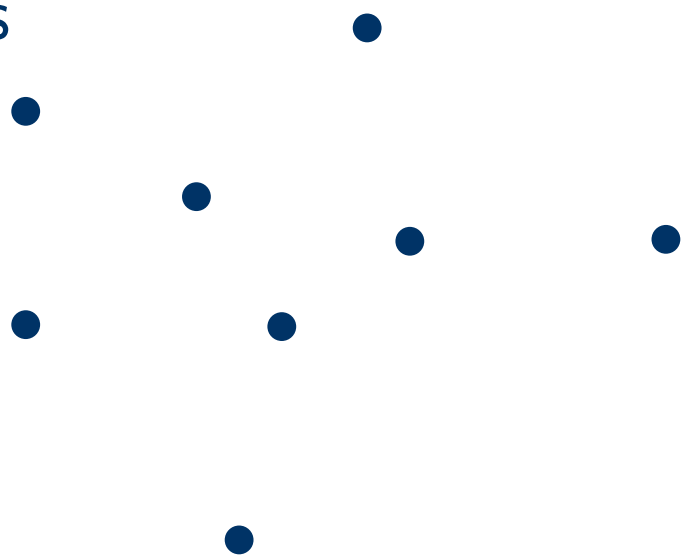


$k = 1$



$k = 2$

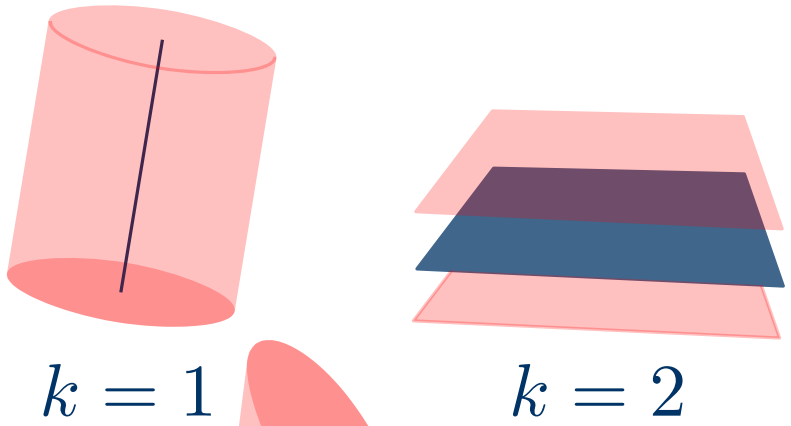
Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points



C_1

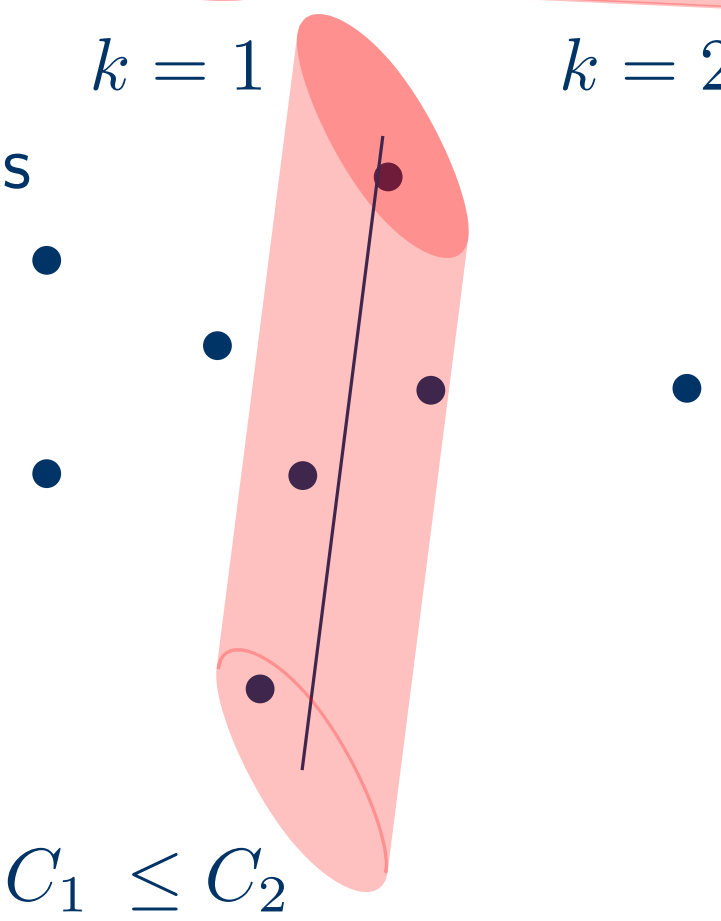
General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



$k = 1$ $k = 2$

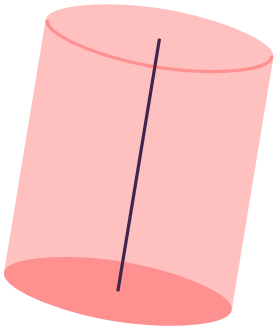
Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points



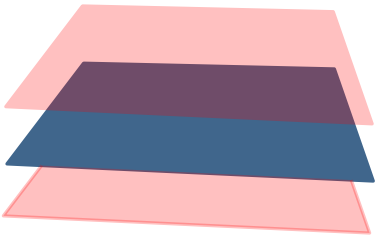
$C_1 \leq C_2$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



$k = 1$



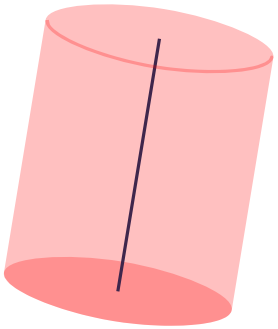
$k = 2$

Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points

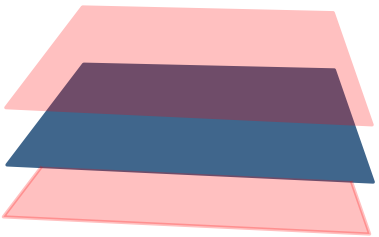
$C_1 \leq C_2$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.

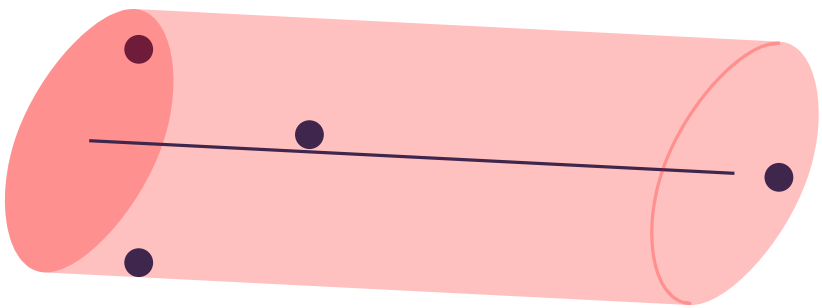


$k = 1$



$k = 2$

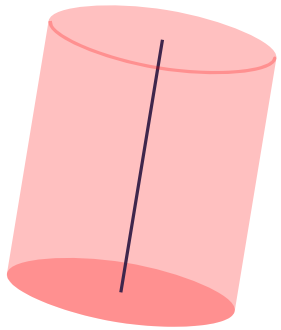
Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points



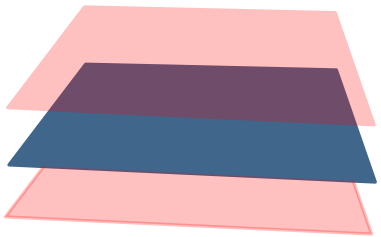
$$C_1 \leq C_2 \leq C_3$$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



$k = 1$



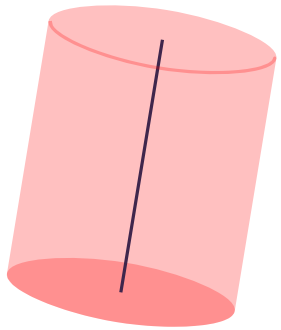
$k = 2$

Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points

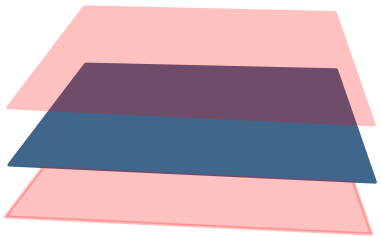
$$C_1 \leq C_2 \leq C_3 \leq \cdots \leq C_{n/m}$$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



$k = 1$



$k = 2$

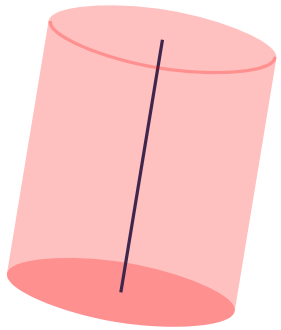
Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points

For each C_i, build cluster structure

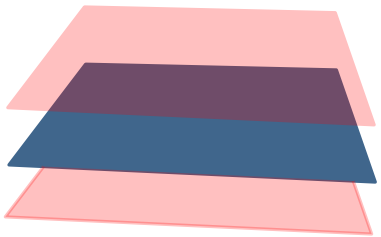
$$C_1 \leq C_2 \leq C_3 \leq \cdots \leq C_{n/m}$$

General Framework (Preprocessing)

Def.: Let K be a k-flat. Then $C := \{p \in \mathbb{R}^d \mid \ d(p, K) \leq \alpha\}$ is a *k-flat cluster* for K with radius α.



$k = 1$



$k = 2$

Iteratively find & remove smallest cluster that contains $m \approx \sqrt{n}$ points

For each C_i, build cluster structure

For suffix $C_i, \ldots, C_{n/m}$, build projection structure

$$C_1 \leq C_2 \leq C_3 \leq \cdots \leq C_{n/m}$$

General Framework (Query)

small parameter $t > 0$, query k-flat F

1. Find a n^t-approximate NN $\hat{p}$ for F, set $\alpha = d(\hat{p}, F)$

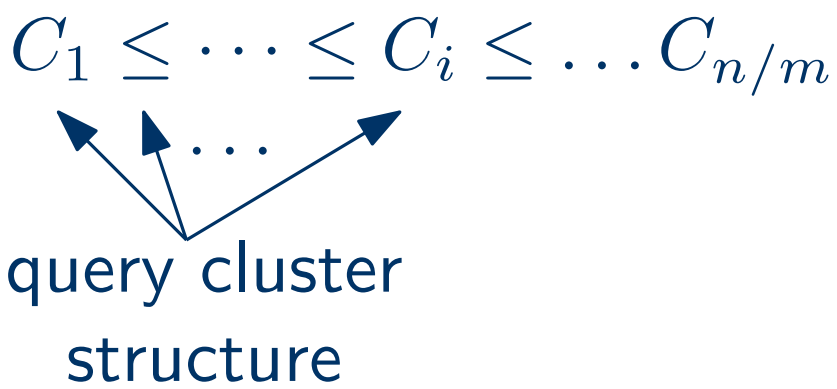
$$C_1 \leq \cdots \leq C_i \leq \ldots C_{n/m}$$

General Framework (Query)

small parameter $t > 0$, query k-flat F

1. Find a n^t-approximate NN $\hat{p}$ for F, set $\alpha = d(\hat{p}, F)$

2. Query cluster structures of clusters with radius $\leq \alpha n^t$ with F



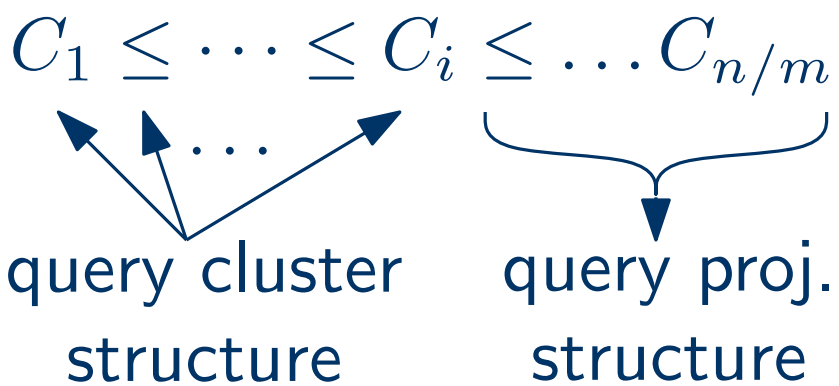
General Framework (Query)

small parameter $t > 0$, query k-flat F

1. Find a n^t-approximate NN $\hat{p}$ for F, set $\alpha = d(\hat{p}, F)$

2. Query cluster structures of clusters with radius $\leq \alpha n^t$ with F

3. Query projection structure of remaining points using F and α

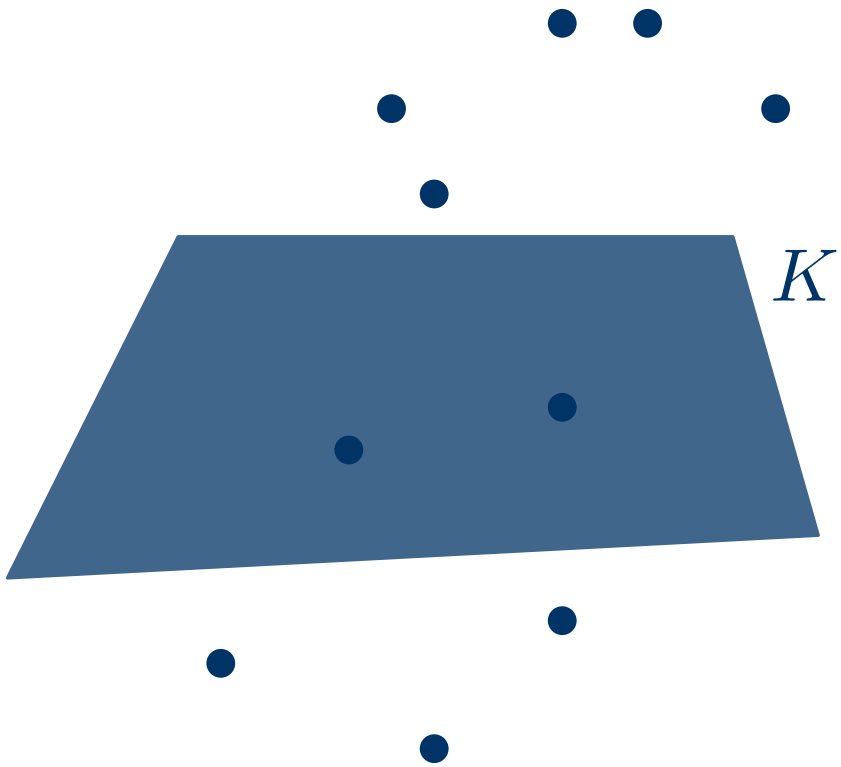


Cluster Structure (Preprocessing)

Given: cluster C as k-flat K and radius αn^t, m points Q in C

Query: k-flat F, find c-apx NN for F in Q

Preprocessing: Build 0-ANN structure $\mathcal{A}$ for Q in $K^\perp$

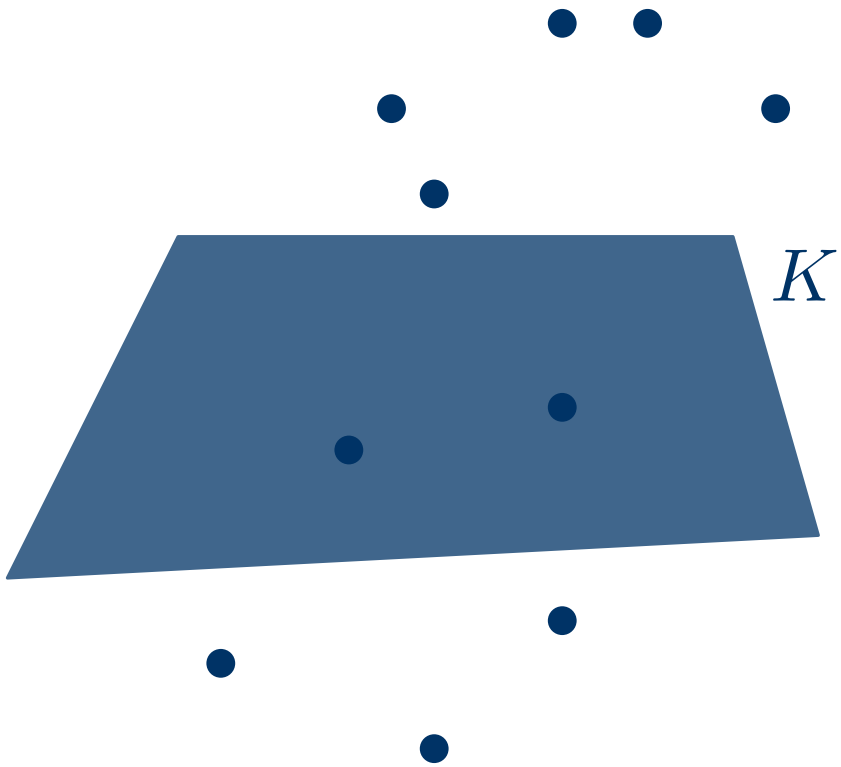


Cluster Structure (Preprocessing)

Given: cluster C as k-flat K and radius αn^t, m points Q in C

Query: k-flat F, find c-apx NN for F in Q

Preprocessing: Build 0-ANN structure $\mathcal{A}$ for Q in $K^\perp$

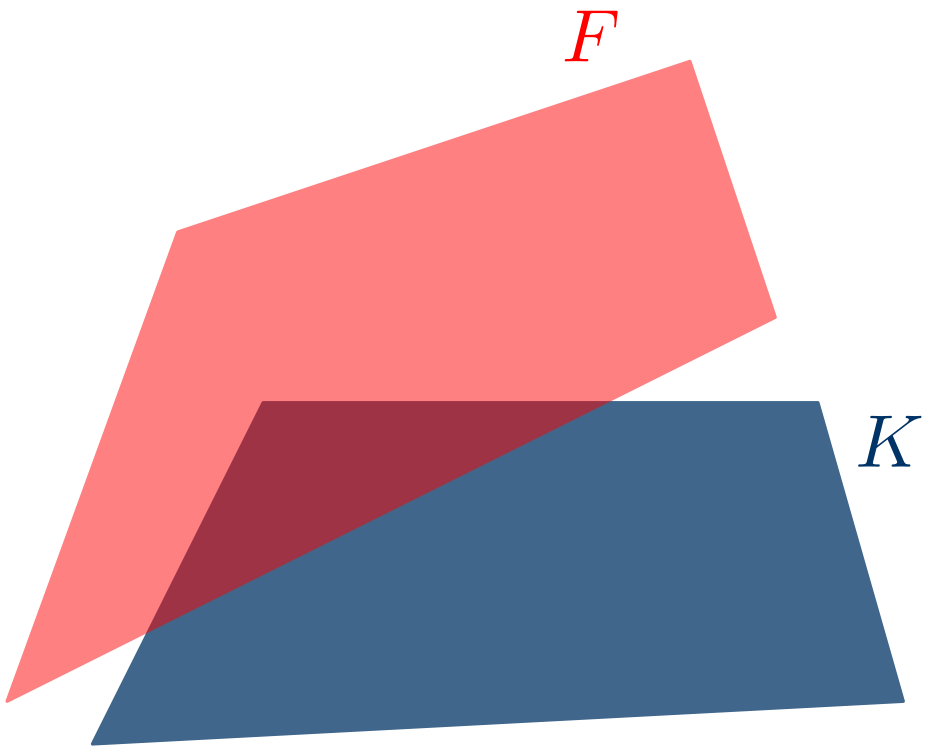


Idea: approximate F by few "patches" $\mathcal{G}$ that are parallel to K, query $\mathcal{A}$ with each $G \in \mathcal{G}$

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

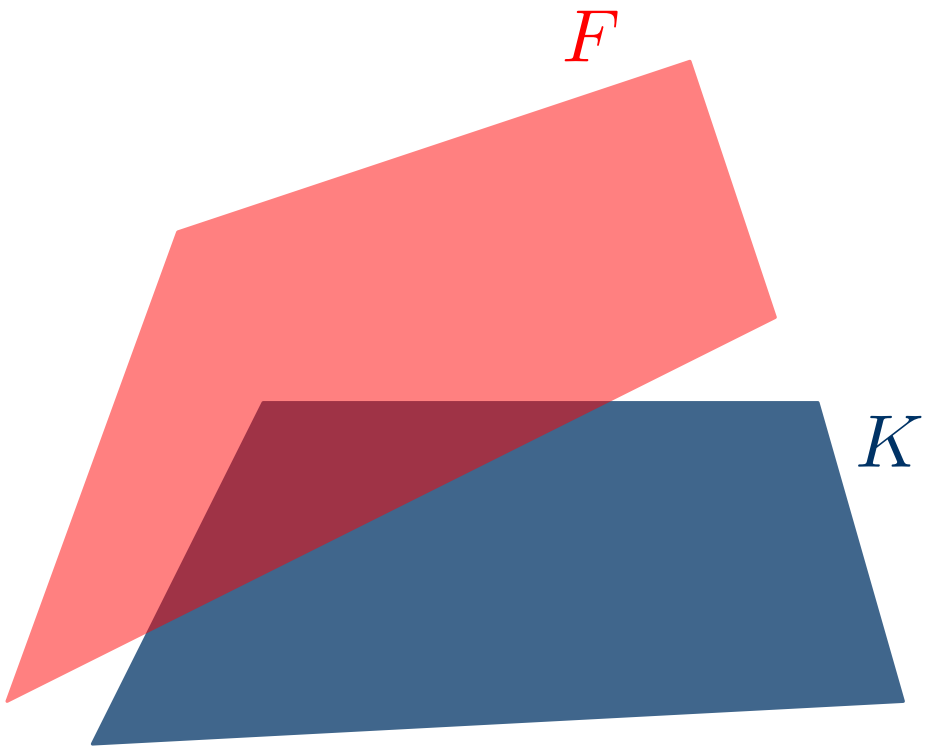


Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U \Sigma V^T$.

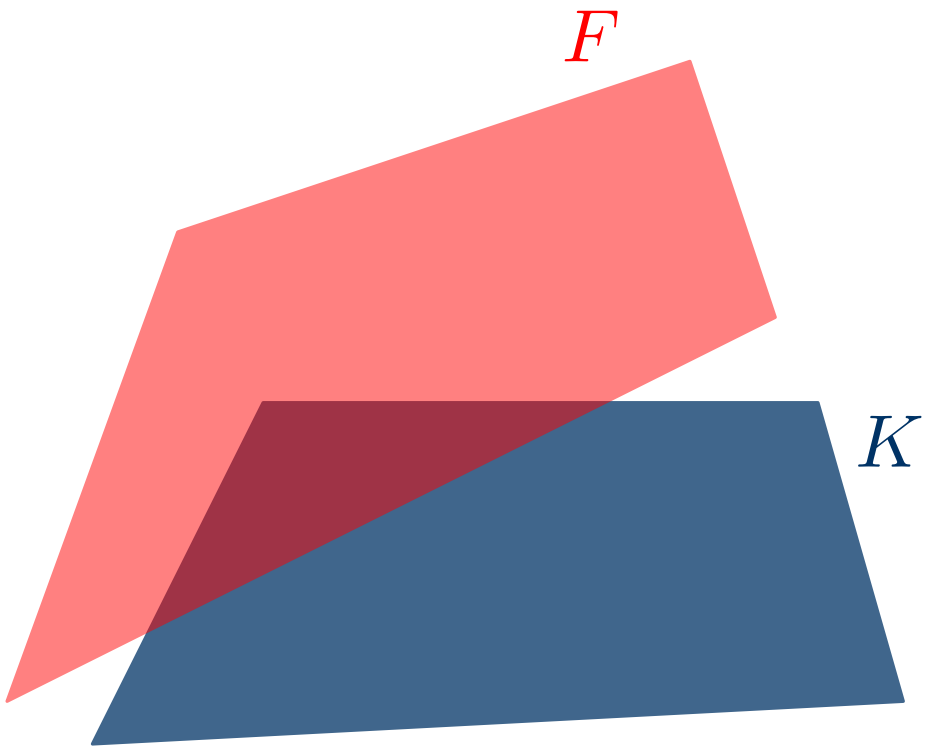


Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U\Sigma V^T$.



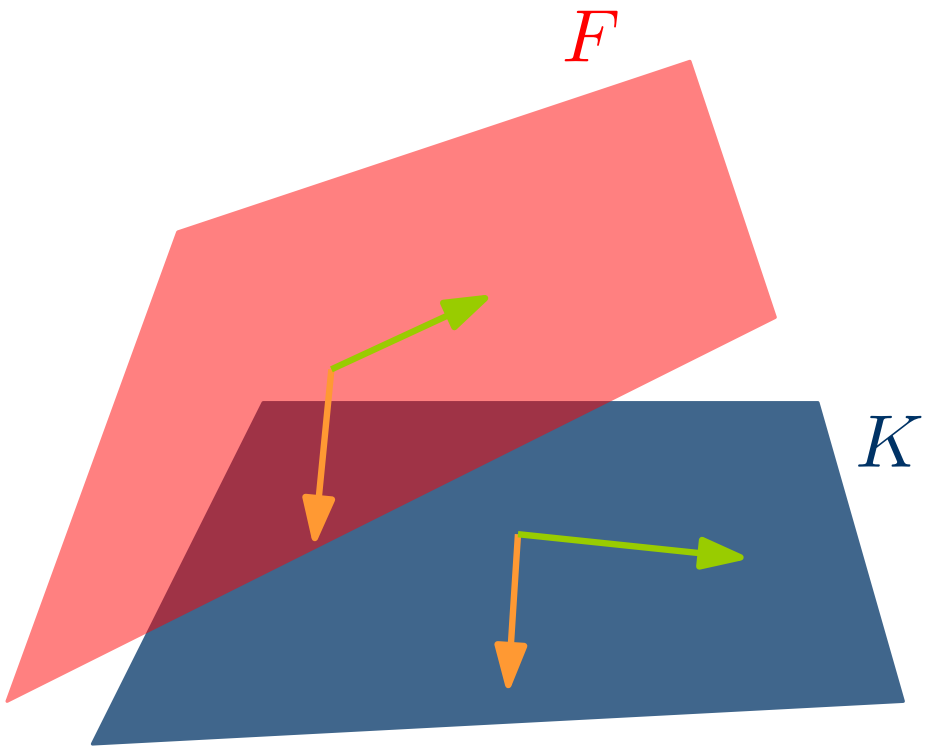
singular values $1 \geq \sigma_1 \geq \cdots \geq \cdots \geq \sigma_l \geq \sigma_{l+1} \geq \cdots \geq \sigma_k \geq 0$

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U\Sigma V^T$.



singular values $1 \geq \underbrace{\sigma_1 \geq \cdots \geq \cdots \geq \sigma_l}_{\text{parallel}} \geq \underbrace{\sigma_{l+1} \geq \cdots \geq \sigma_k}_{\text{orthogonal}} \geq 0$

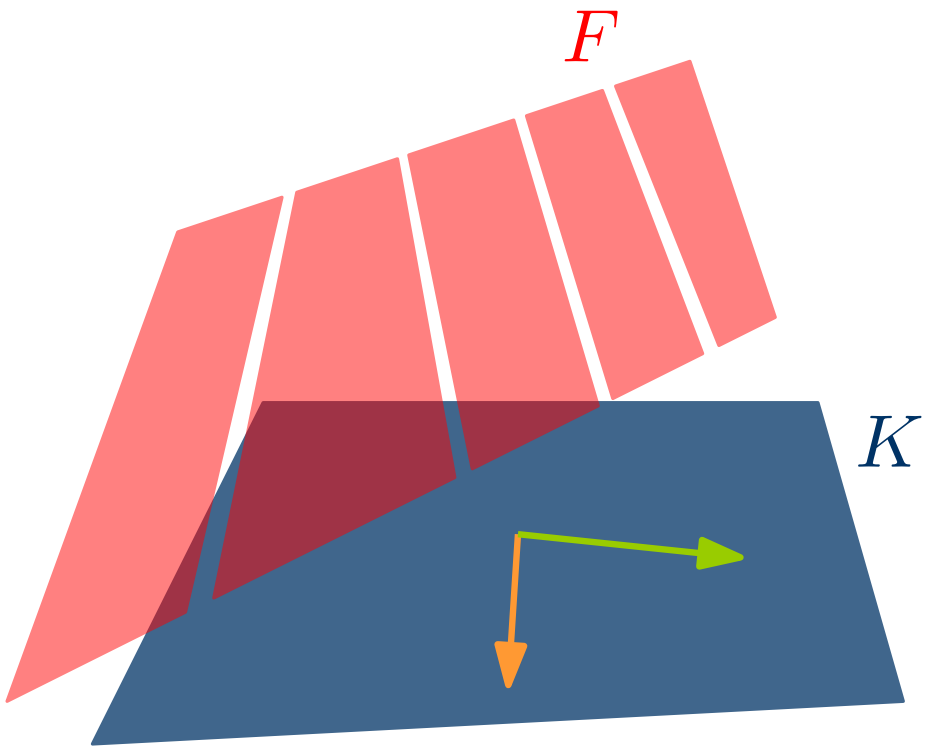
$\sigma_l \geq \sqrt{1-\varepsilon} > \sigma_{l+1}$

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U \Sigma V^T$.



singular values $1 \geq \underbrace{\sigma_1 \geq \cdots \geq \cdots \geq \sigma_l}_{\text{parallel}} \geq \underbrace{\sigma_{l+1} \geq \cdots \geq \sigma_k}_{\text{orthogonal}} \geq 0$

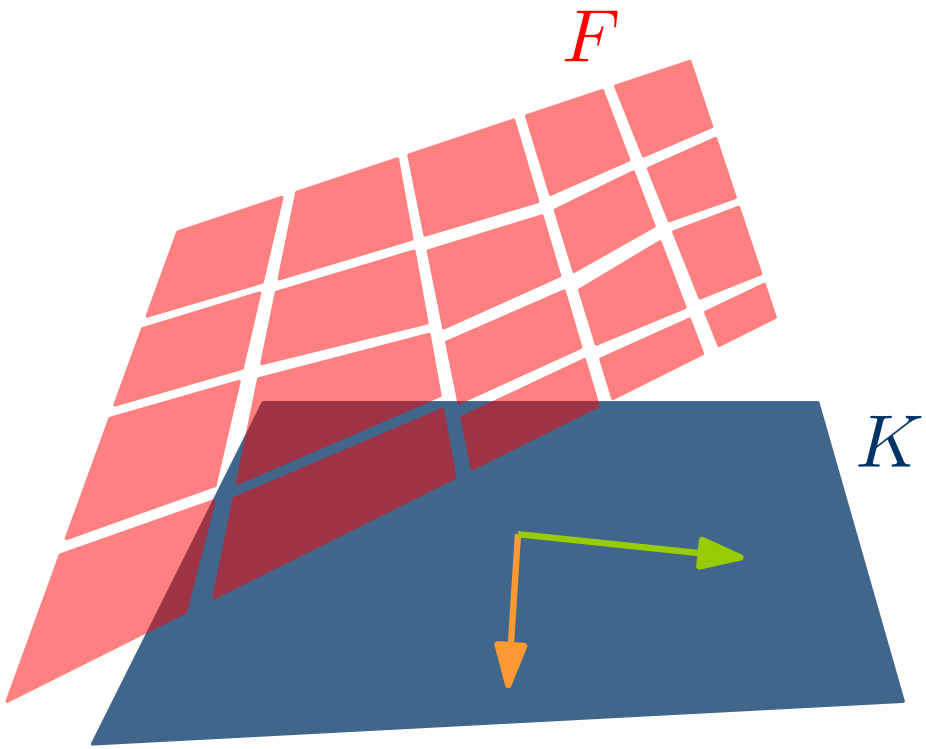
discretize F by
l-dimensional slabs along
the parallel directions

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U\Sigma V^T$.



singular values $1 \geq \underbrace{\sigma_1 \geq \cdots \geq \cdots \geq \sigma_l}_{\text{parallel}} \geq \underbrace{\sigma_{l+1} \geq \cdots \geq \sigma_k}_{\text{orthogonal}} \geq 0$

discretize F by l-dimensional slabs along the parallel directions

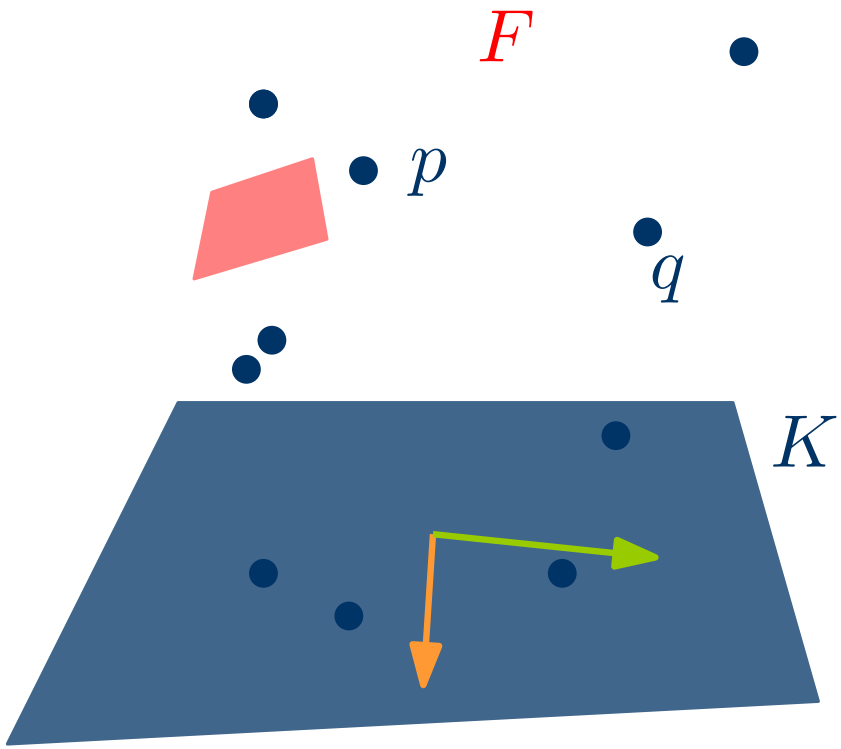
Lemma: We can find $O((n^{2t} k \varepsilon^{-2})^l)$ patches $\mathcal{G}$, s.t. $\exists G \in \mathcal{G}$ with $d(G, Q) \leq (1 + \varepsilon) d(F, Q)$.

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U\Sigma V^T$.



singular values $1 \geq \underbrace{\sigma_1 \geq \cdots \geq \cdots \geq \sigma_l}_{\text{parallel}} \geq \underbrace{\sigma_{l+1} \geq \cdots \geq \sigma_k}_{\text{orthogonal}} \geq 0$

discretize F by l-dimensional slabs along the parallel directions

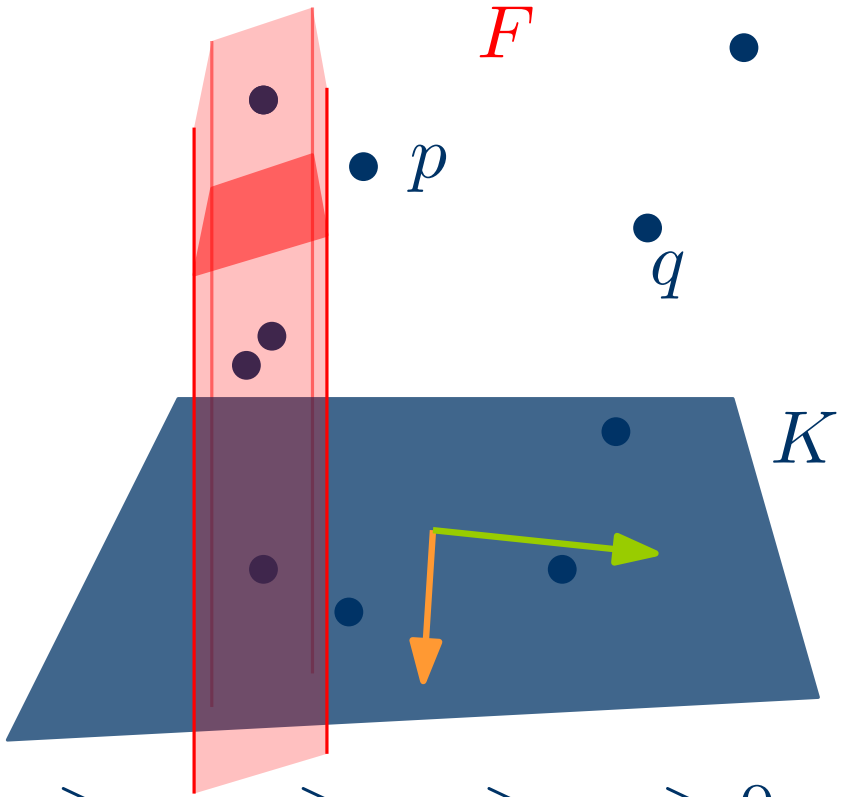
Lemma: We can find $O((n^{2t} k \varepsilon^{-2})^l)$ patches $\mathcal{G}$, s.t. $\exists G \in \mathcal{G}$ with $d(G, Q) \leq (1 + \varepsilon) d(F, Q)$.

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U \Sigma V^T$.



singular values $1 \geq \underbrace{\sigma_1 \geq \cdots \geq \cdots \geq \sigma_l}_{\text{parallel}} \geq \underbrace{\sigma_{l+1} \geq \cdots \geq \sigma_k}_{\text{orthogonal}} \geq 0$

discretize F by l-dimensional slabs along the parallel directions

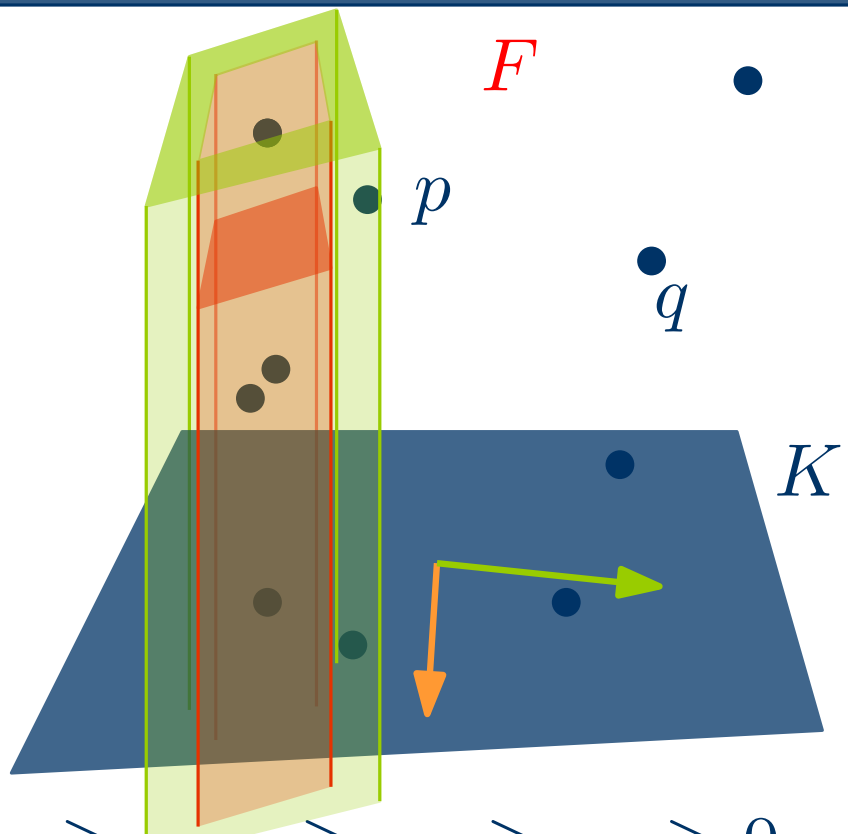
Lemma: We can find $O((n^{2t} k \varepsilon^{-2})^l)$ patches $\mathcal{G}$, s.t. $\exists G \in \mathcal{G}$ with $d(G, Q) \leq (1 + \varepsilon) d(F, Q)$.

Cluster Structure (Query)

cluster flat $K : v \mapsto Av + a$
query flat $F : v \mapsto Bv + b$
with $A, B \in \mathbb{R}^{d \times k}$ and $a, b \in \mathbb{R}^d$

$\varepsilon = 1/100 \log n$

Compute $M = A^T B$ and the *singular value decomposition* $M = U\Sigma V^T$.



singular values $1 \geq \underbrace{\sigma_1 \geq \cdots \geq \cdots \geq \sigma_l}_{\text{parallel}} \geq \underbrace{\sigma_{l+1} \geq \cdots \geq \sigma_k}_{\text{orthogonal}} \geq 0$

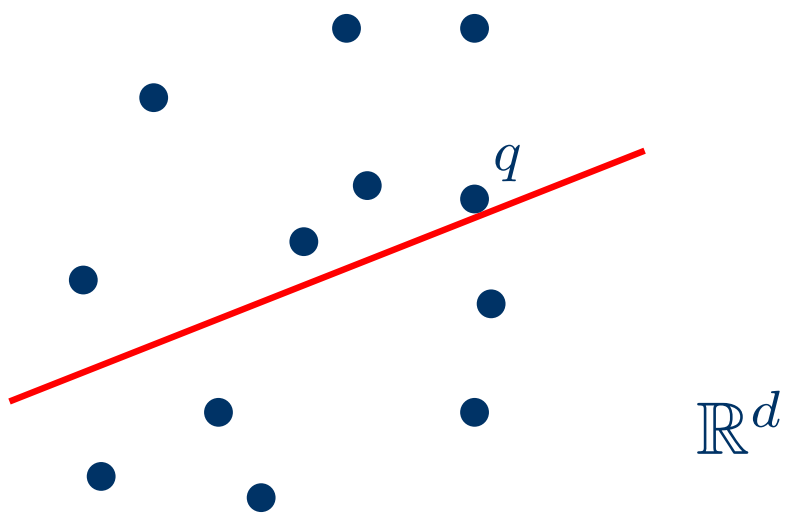
discretize F by l-dimensional slabs along the parallel directions

Lemma: We can find $O((n^{2t} k \varepsilon^{-2})^l)$ patches $\mathcal{G}$, s.t. $\exists G \in \mathcal{G}$ with $d(G, Q) \leq (1 + \varepsilon) d(F, Q)$.

Projection Structure (Idea)

Given: n points $P \subset \mathbb{R}^d$ s.t. no cluster with radius $\leq \alpha n^t$ contains more that m points

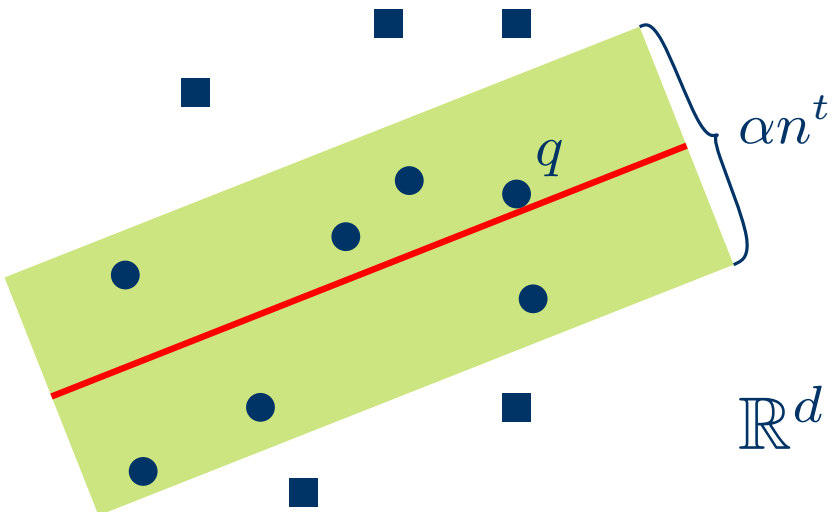
Query: k-flat F, find NN $q \in Q$



Projection Structure (Idea)

Given: n points $P \subset \mathbb{R}^d$ s.t. no cluster with radius $\leq \alpha n^t$ contains more that m points

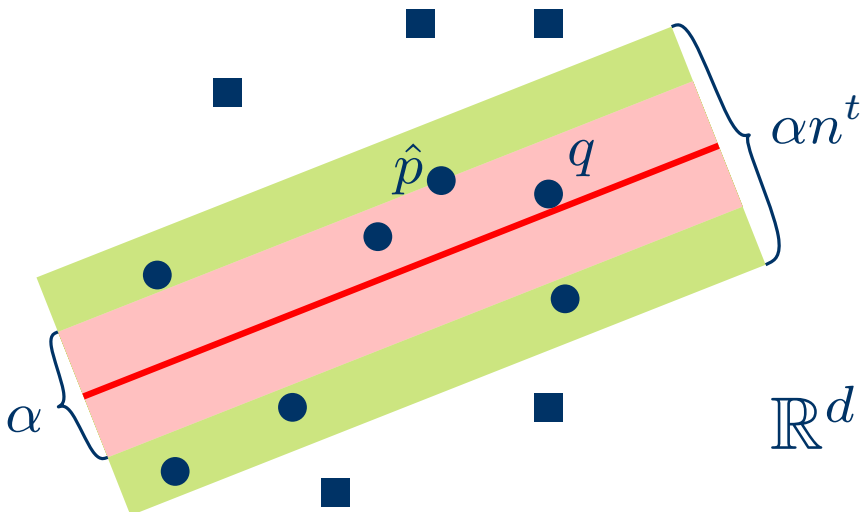
Query: k-flat F, find NN $q \in Q$



Projection Structure (Idea)

Given: n points $P \subset \mathbb{R}^d$ s.t. no cluster with radius $\leq \alpha n^t$ contains more that m points

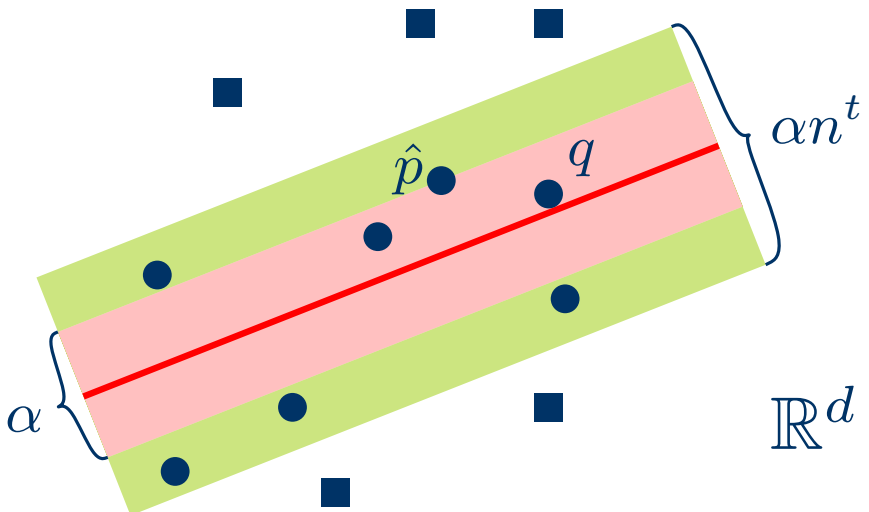
Query: k-flat F, find NN $q \in Q$



Projection Structure (Idea)

Given: n points $P \subset \mathbb{R}^d$ s.t. no cluster with radius $\leq \alpha n^t$ contains more that m points

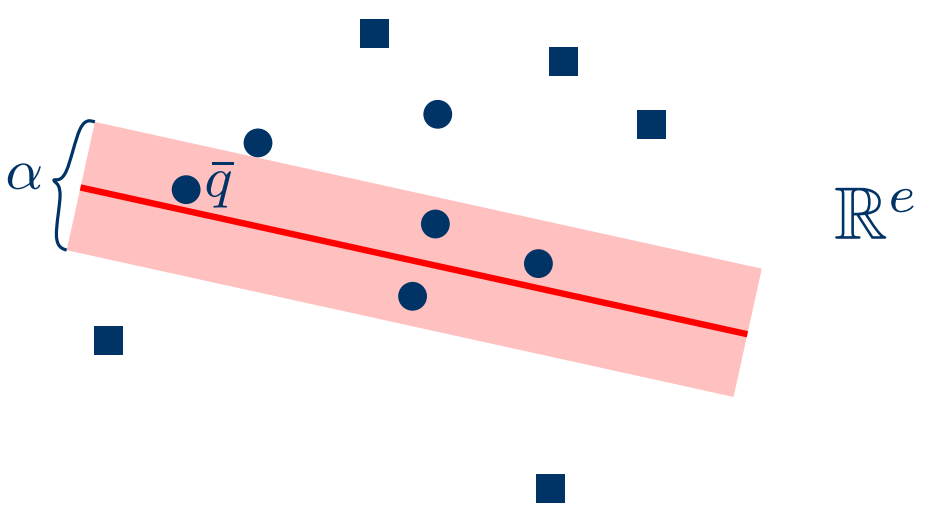
Query: k-flat F, find NN $q \in Q$



random proj. onto $e = O(1/t)$-dim space by JL-Lemma

Given: $\bar{P} \subset \mathbb{R}^e$

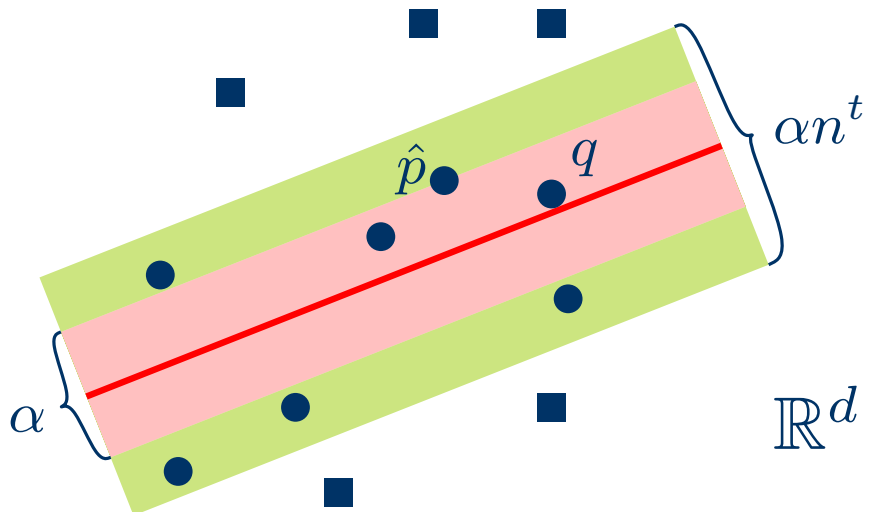
Query: k-flat $\bar{F} \subset \mathbb{R}^e$, $\alpha \in \mathbb{R}$, find $R = \{p \in \bar{P} \mid d(\bar{F}, \bar{p}) \leq \alpha\}$



Projection Structure (Idea)

Given: n points $P \subset \mathbb{R}^d$ s.t. no cluster with radius $\leq \alpha n^t$ contains more that m points

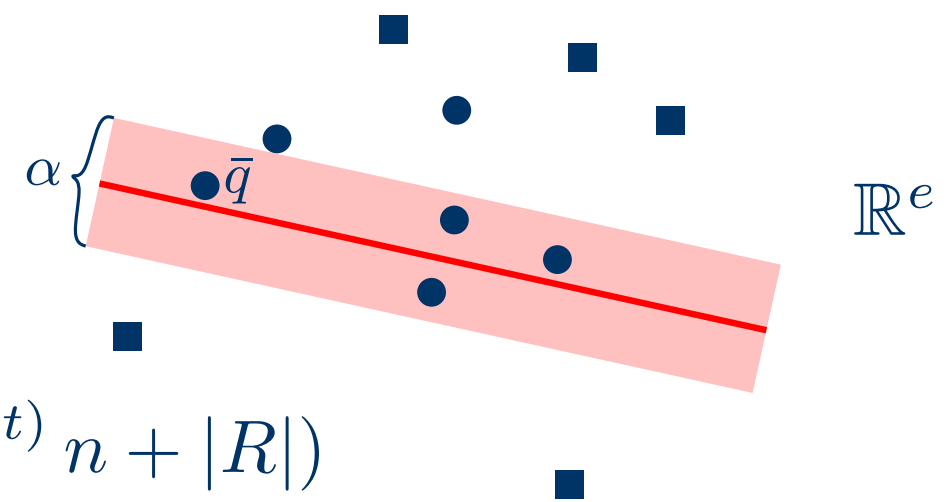
Query: k-flat F, find NN $q \in Q$



random proj. onto $e = O(1/t)$-dim space by JL-Lemma

Given: $\bar{P} \subset \mathbb{R}^e$

Query: k-flat $\bar{F} \subset \mathbb{R}^e$, $\alpha \in \mathbb{R}$, find $R = \{p \in \bar{P} \mid d(\bar{F}, \bar{p}) \leq \alpha\}$



space $O(n \log^{O(1/t)} n)$

query time $O(n^{k/(k+1)} \log^{O(1/t)} n + |R|)$

Summary

Theorem 1: k-ANN can be solved using **space** $O(d^{O(1)} n^{1+k\sigma/(k+1-\rho)} + n \log^{O(1/t)} n)$ and **query time** $O(d^{O(1)} n^{k/(k+1-\rho)+t})$.

random proj. onto $O(1/t)$ dim.

Theorem 2 (Cluster Structure): For m points contained in a k-flat cluster with radius αn^t, we can solve k-ANN with **space** $O(m^{1+\sigma} + d \log^2 m)$ and **query time** $O((n^{2t}k^2)^{k+1} m^{1-1/k+\rho/k})$.

Theorem 3 (Near Neighbors): We can solve approximate *near* neighbor search for k-flats with **space** $O(n \log^{O(1/t} n)$ and **query time** $O(n^{k/(k+1)} \log^{O(1/t)} n + |R|)$.

Summary

Theorem 1: k-ANN can be solved using **space** $O(d^{O(1)} n^{1+k\sigma/(k+1-\rho)} + n \log^{O(1/t)} n)$ and **query time** $O(d^{O(1)} n^{k/(k+1-\rho)+t})$.

random proj. onto $O(1/t)$ dim.

Theorem 2 (Cluster Structure): For m points contained in a k-flat cluster with radius αn^t, we can solve k-ANN with **space** $O(m^{1+\sigma} + d \log^2 m)$ and **query time** $O((n^{2t}k^2)^{k+1} m^{1-1/k+\rho/k})$.

Theorem 3 (Near Neighbors): We can solve approximate *near* neighbor search for k-flats with **space** $O(n \log^{O(1/t} n)$ and **query time** $O(n^{k/(k+1)} \log^{O(1/t)} n + |R|)$.

??? Questions ???