# LoRANN: Low-Rank Matrix Factorization for Approximate Nearest Neighbor Search

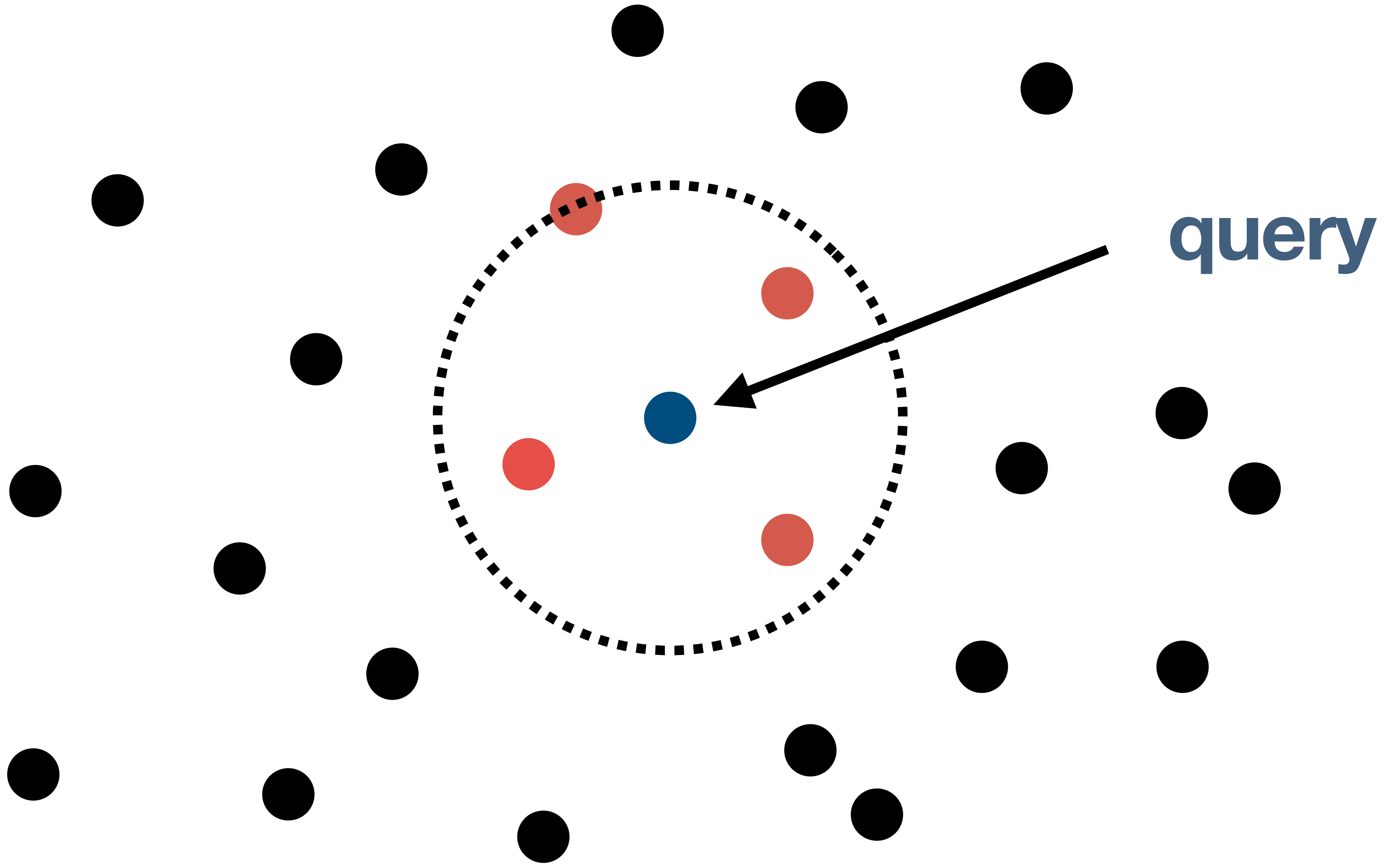
**Elias Jääsaari *, Ville Hyvönen ⌘, Teemu Roos ***

* University of Helsinki ⌘ Aalto University

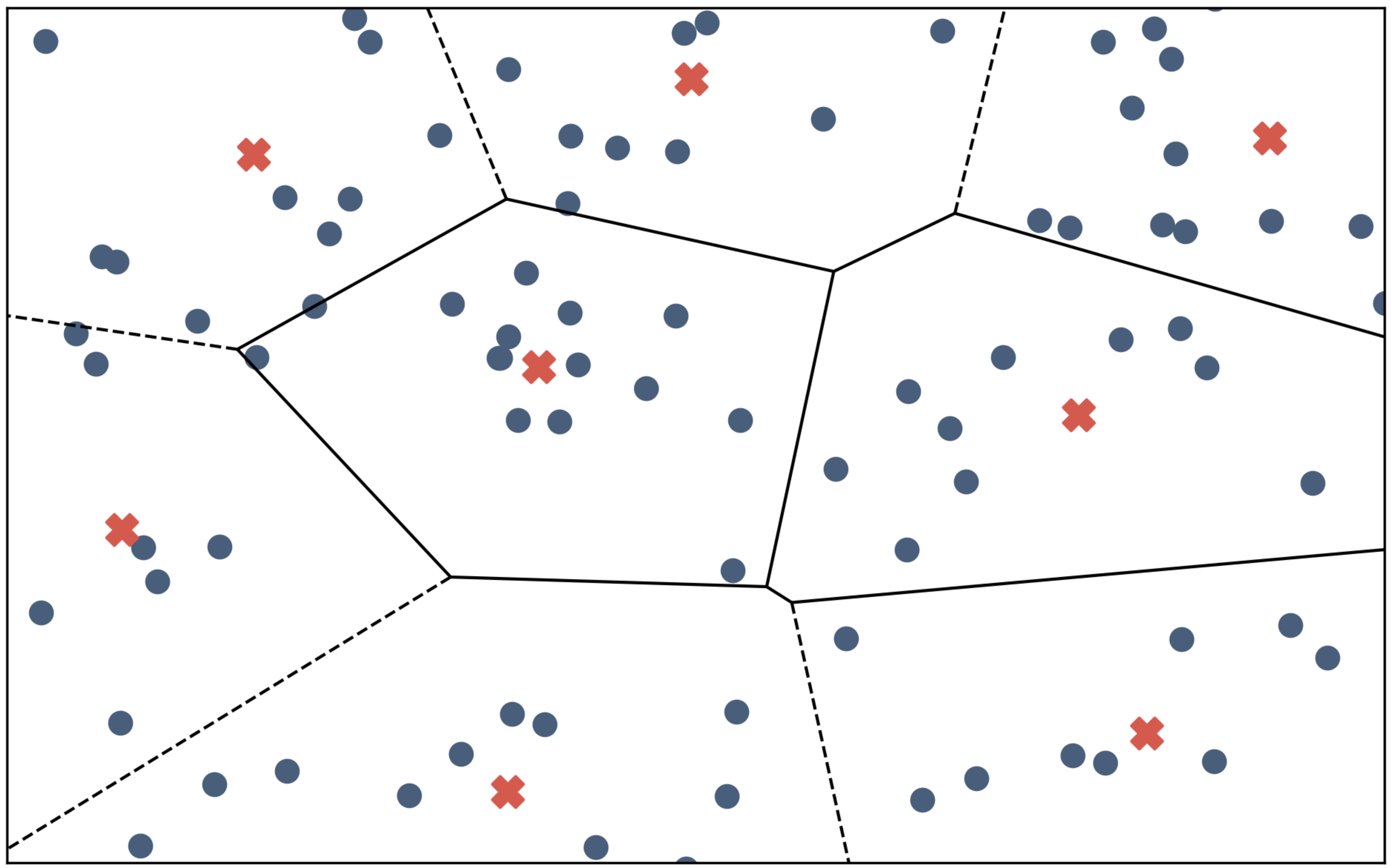
**NeurIPS 2024**

# Approximate nearest neighbor search

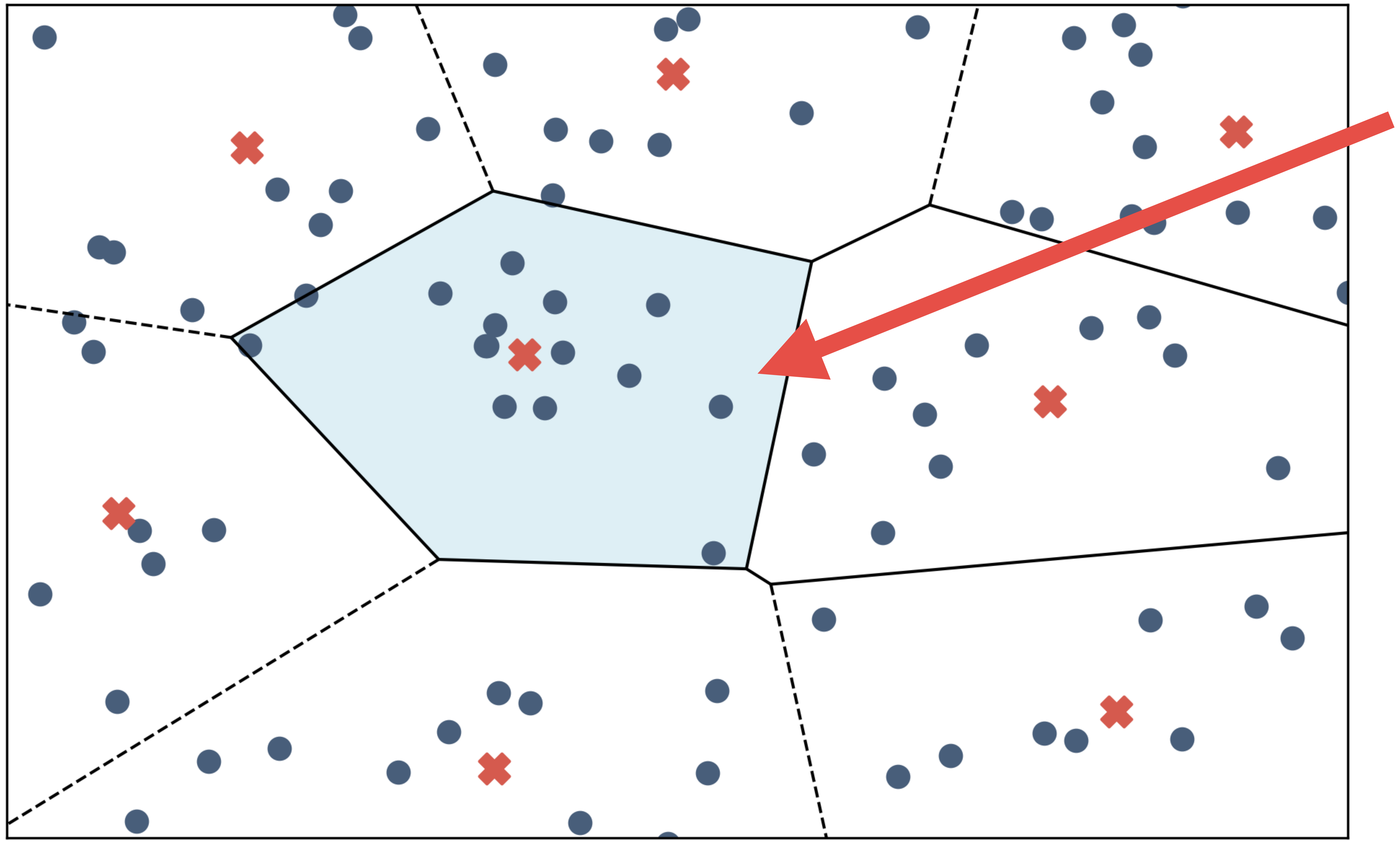
- **Nearest neighbor search** is a key component in many machine learning pipelines
- **Approximate** search methods can be used to speed up the search



# Clustering-based indexes

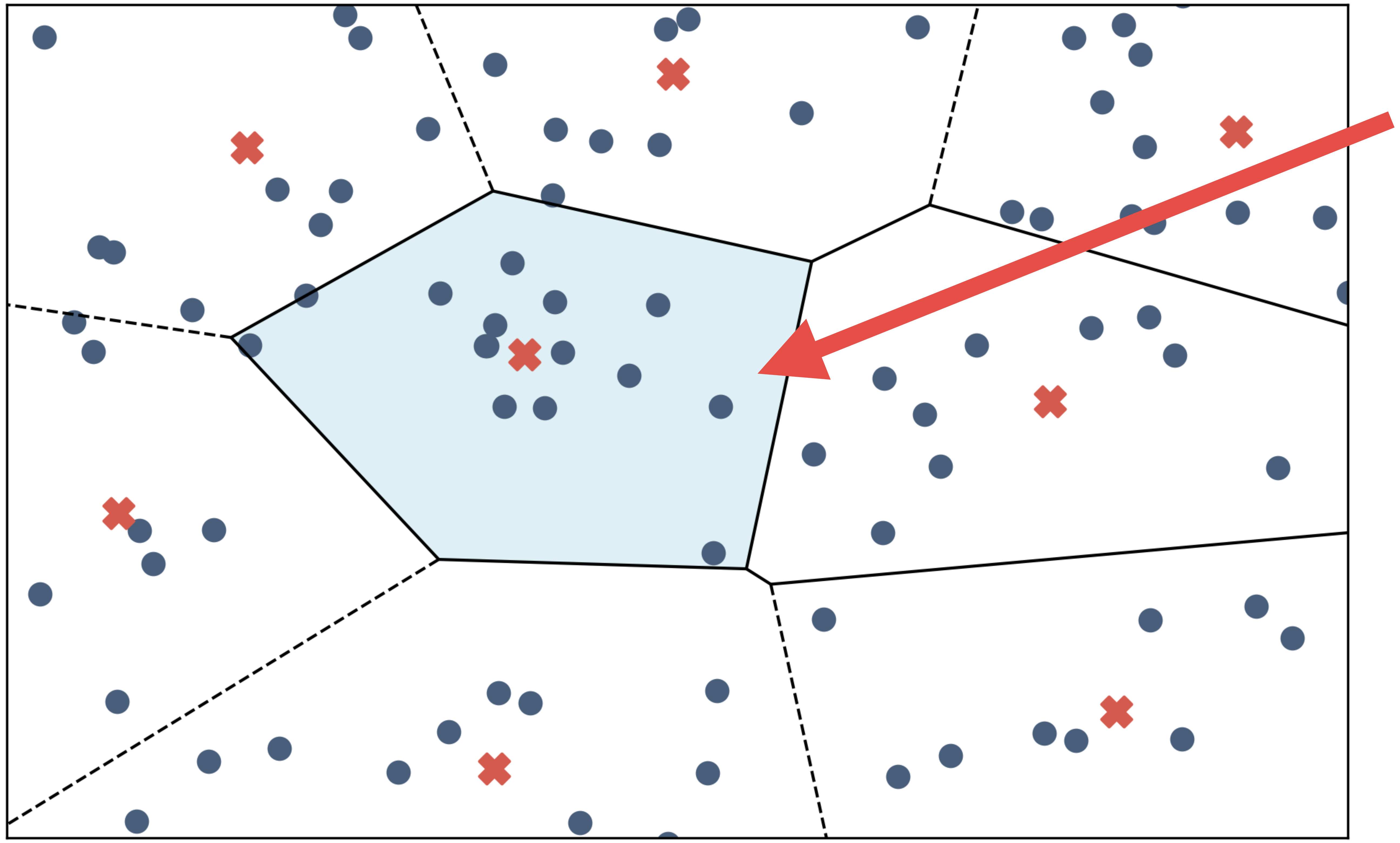


# Clustering-based indexes



Estimate similarities using a **score computation** function

# Clustering-based indexes

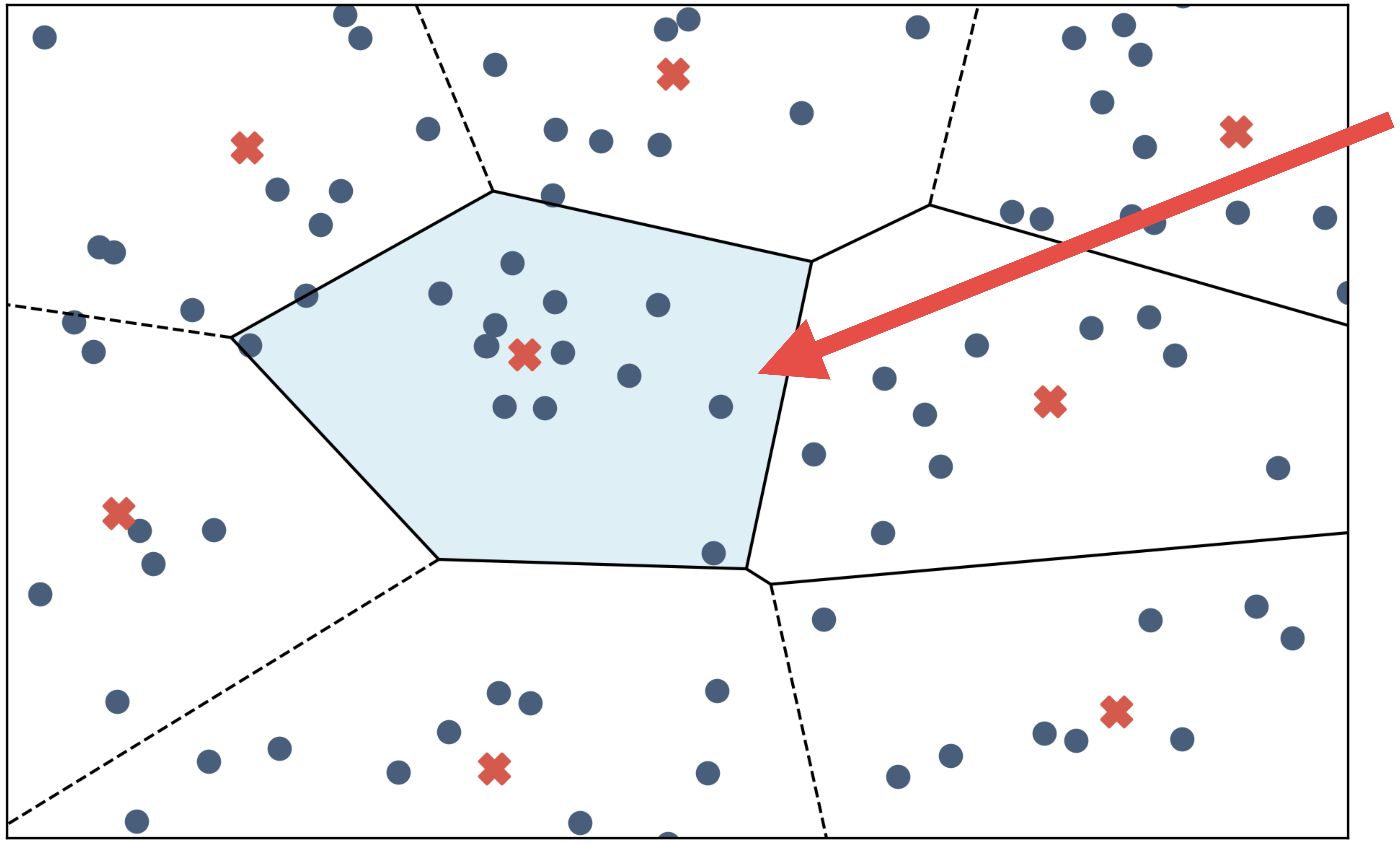


Estimate similarities using a **score computation** function

Compared to graph-based indexes:

🚫 Latency (at high recall rates)
✅ Memory consumption
✅ Index construction time

# Clustering-based indexes



Estimate similarities using a **score computation** function

Compared to graph-based indexes:

🚫 Latency (at high recall rates)
✅ Memory consumption
✅ Index construction time

**Can we make clustering-based indexes as fast as graph-based indexes?**

# Reduced-rank regression

- **Score computation** in clustering-based indexes requires us to estimate the inner products $\mathbf{q}^T\mathbf{C}^T$ between a query point $\mathbf{q}$ and a set of cluster points $\mathbf{C}$
- Our key idea is to formulate score computation as a multivariate regression problem. We estimate the inner products as $\mathbf{q}^T\hat{\beta}$, where $\hat{\beta}$ is a **low-rank** matrix obtained using **reduced-rank regression** (RRR)

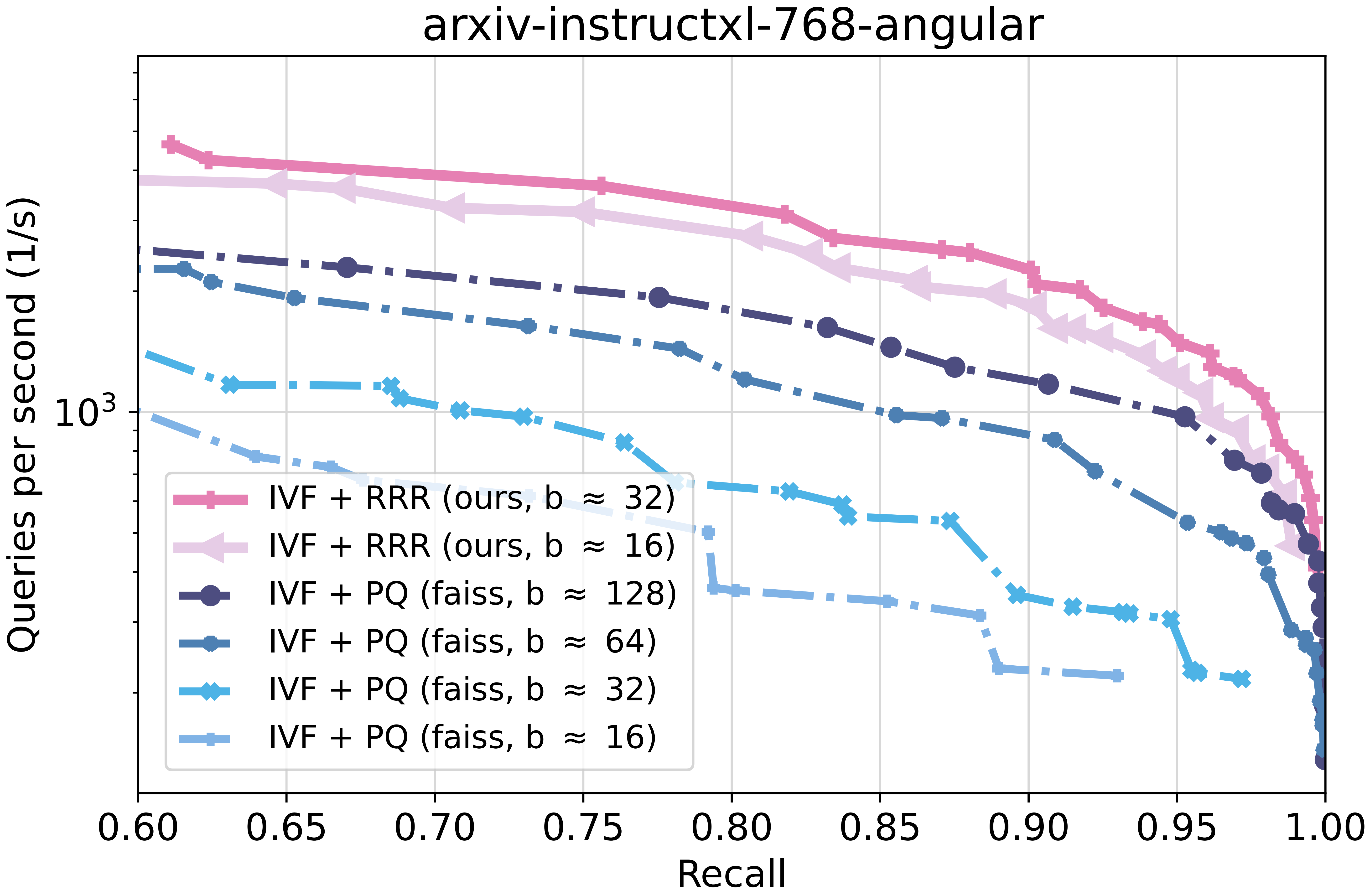
$$\hat{\beta} = \underset{\beta:\text{rank}(\beta)\leq r}{\arg\min} \ \|\mathbf{X}\mathbf{C}^T - \mathbf{X}\beta\|_F^2$$

for a set of training queries $\mathbf{X}$.

- We factor $\hat{\beta} := \mathbf{A}\mathbf{B}$ and apply **8-bit integer quantization** to $\mathbf{q}$, $\mathbf{A}$, and $\mathbf{B}$

# Results

When comparing reduced-rank regression (RRR) against product quantization (PQ) such that **bytes per vector $b$** is the same, **RRR is superior to PQ**



# LoRANN

- We also introduce **LoRANN**, an ANN library leveraging RRR
- LoRANN is competitive with the state-of-the-art graph-based libraries but with **fast indexing** and **tiny memory usage**
- LoRANN achieves **state-of-the-art GPU query latency**

**https://github.com/ejaasaari/lorann**