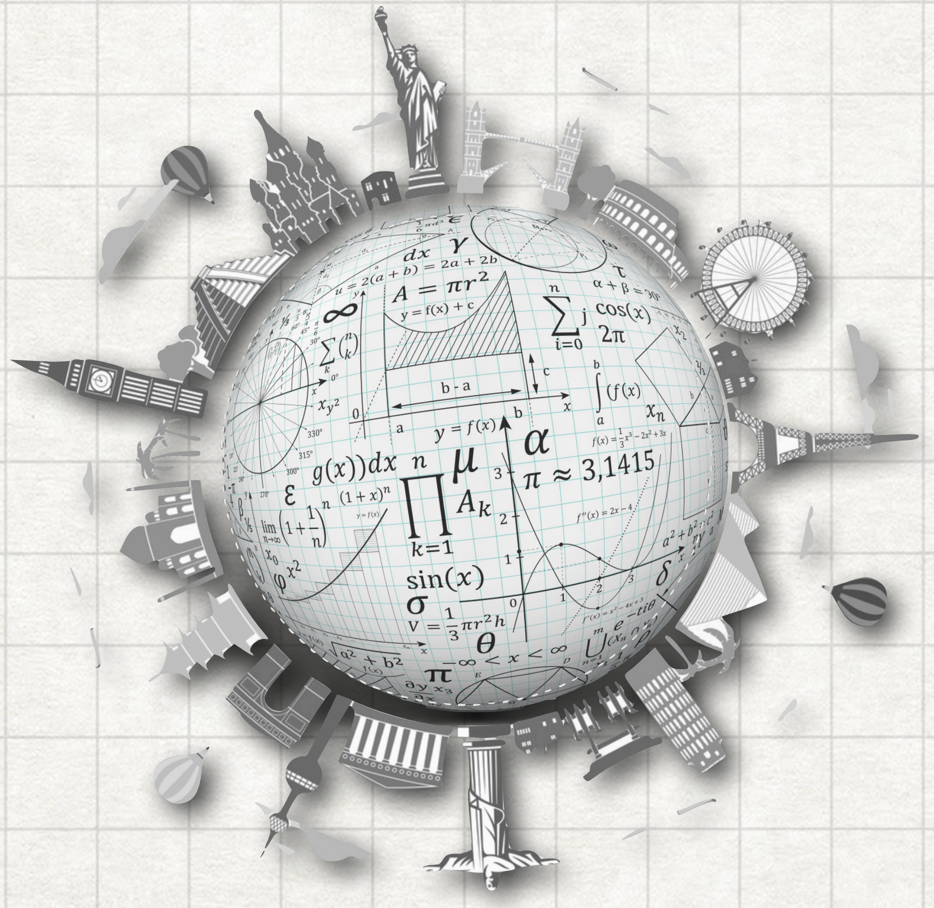
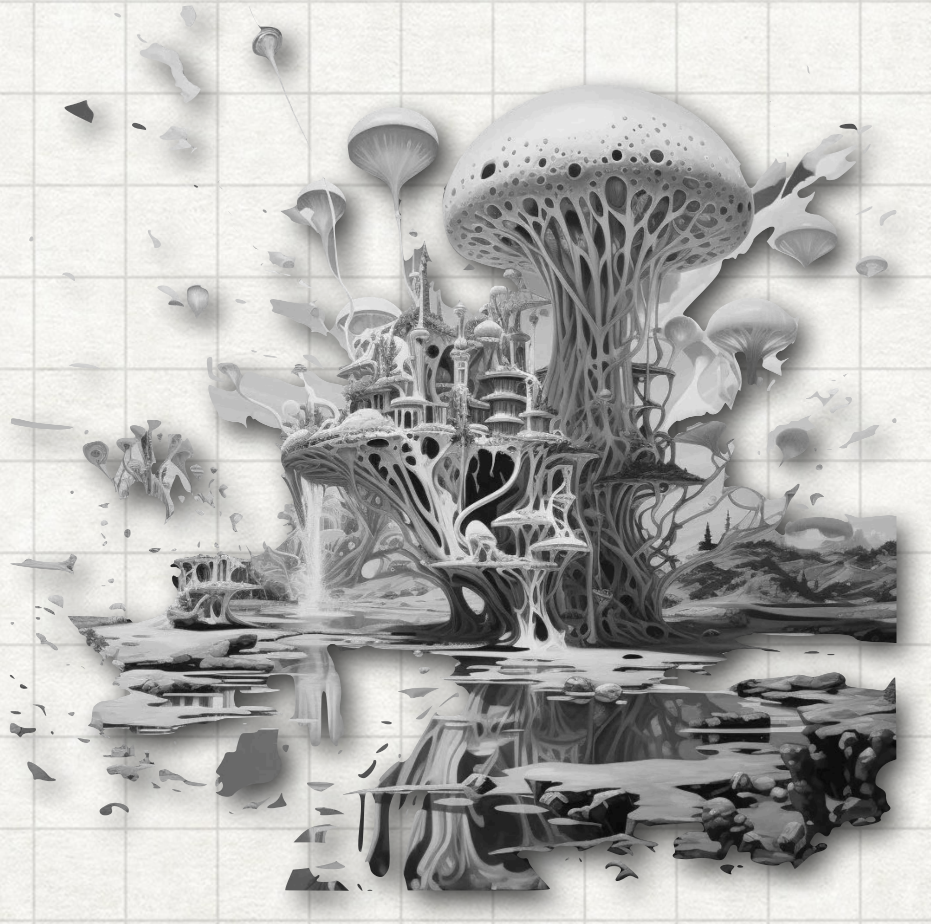
Sebastian Bruch
Pinecone

Information Retrieval Needs More Theoreticians

A Talk in Three Acts





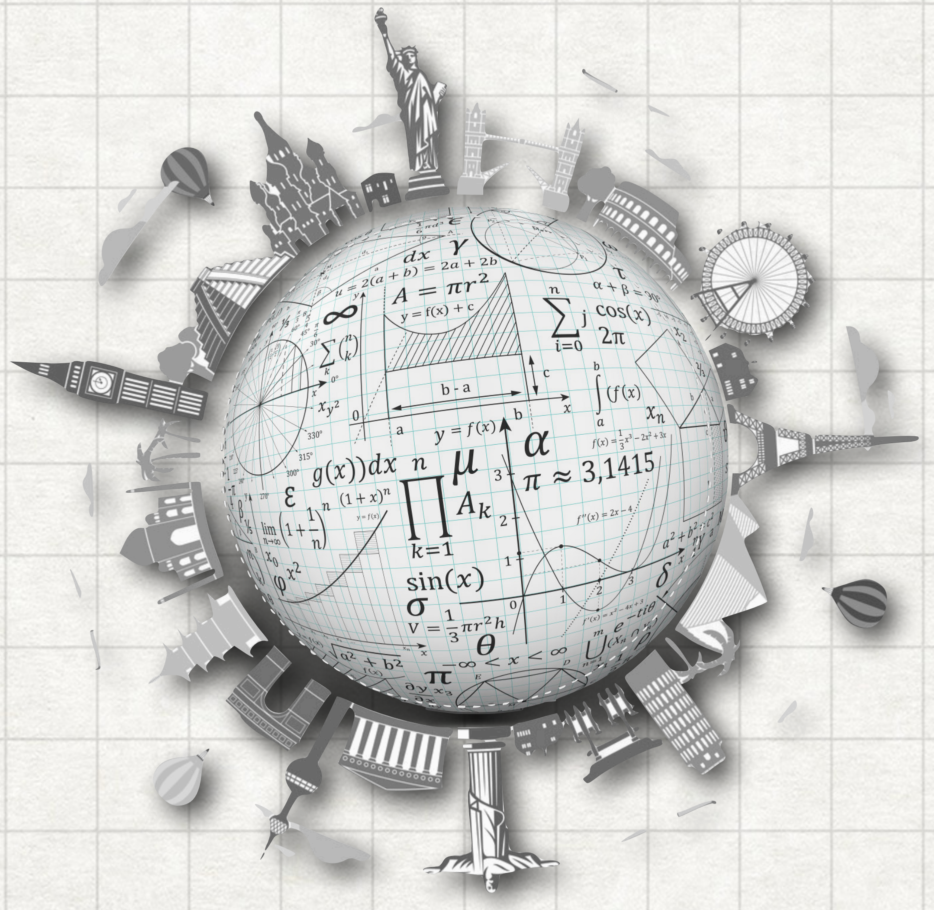
Trigger Warning

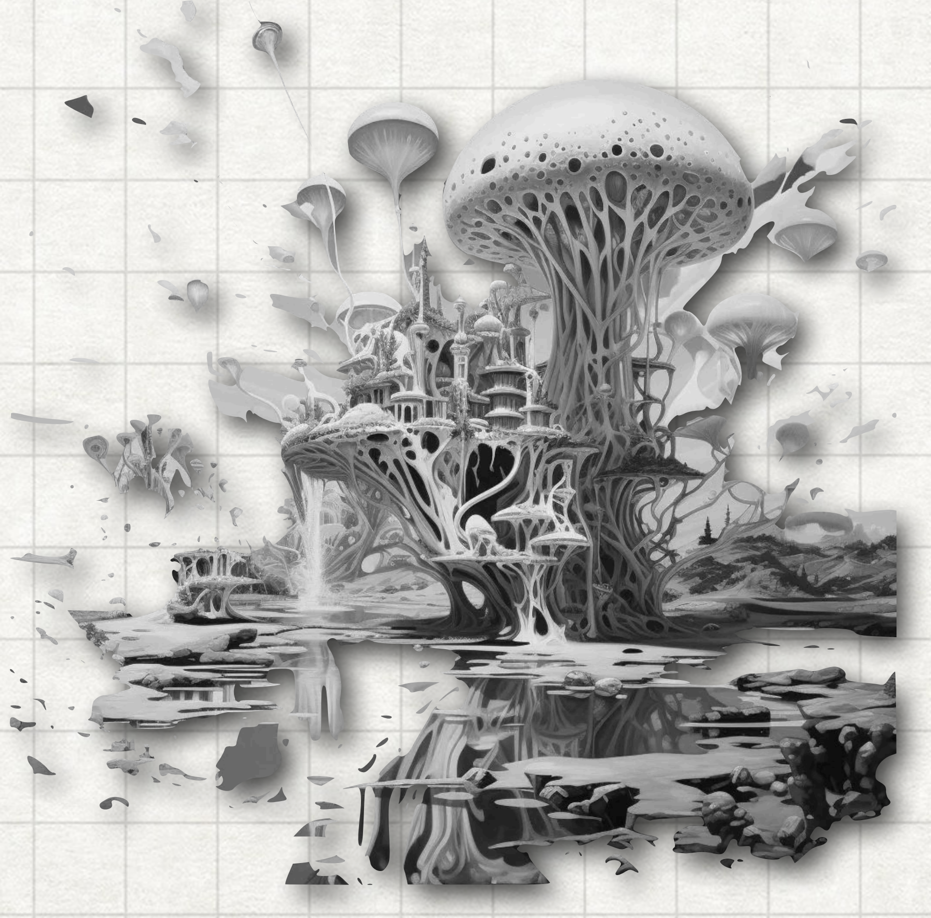
Numerical Linear Algebra

Prob. Data Structures

Mention of Neural Networks

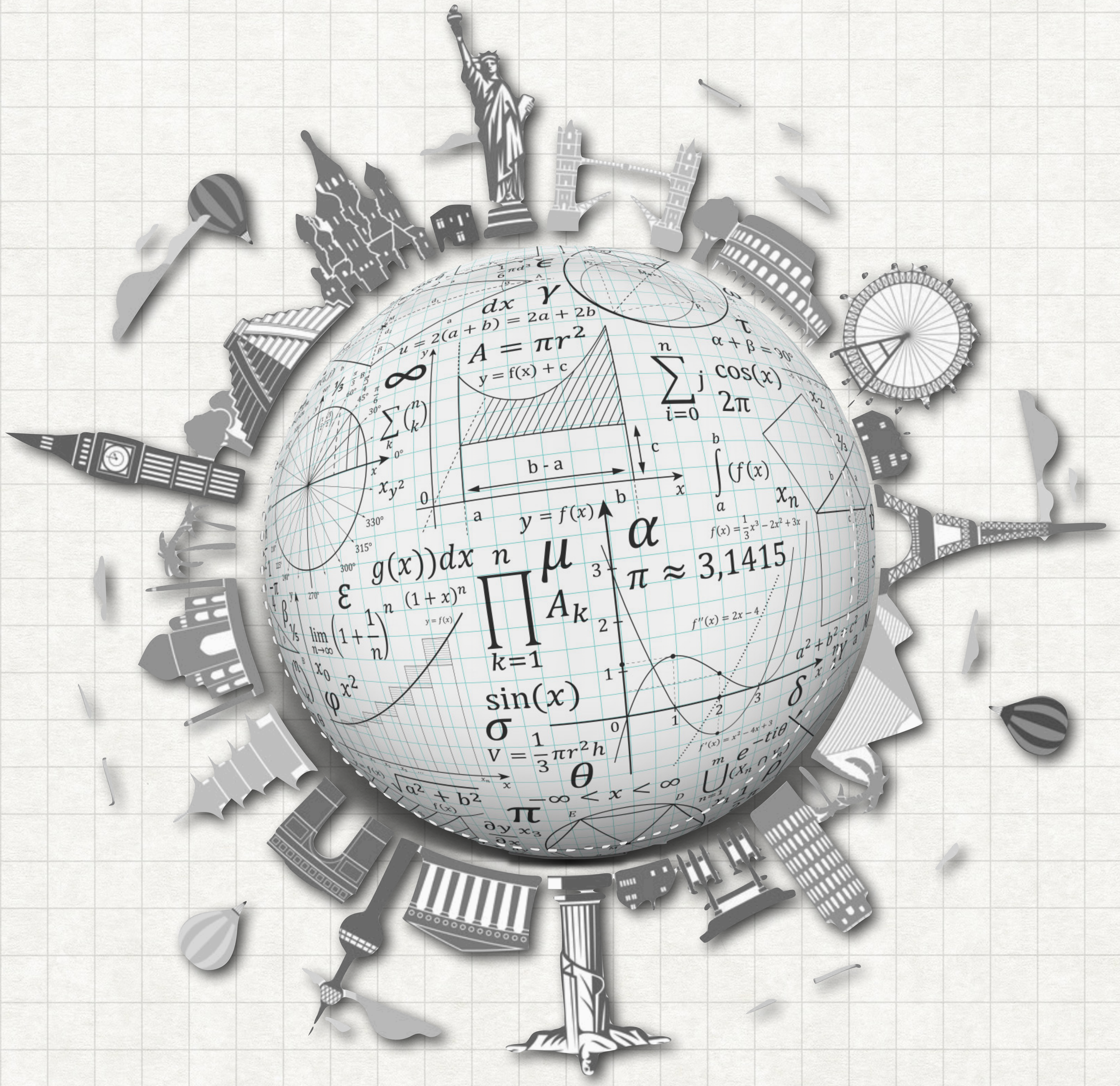
A Talk in Three Acts





Act I: Math

Standing on the Shoulders of Theoretical Giants



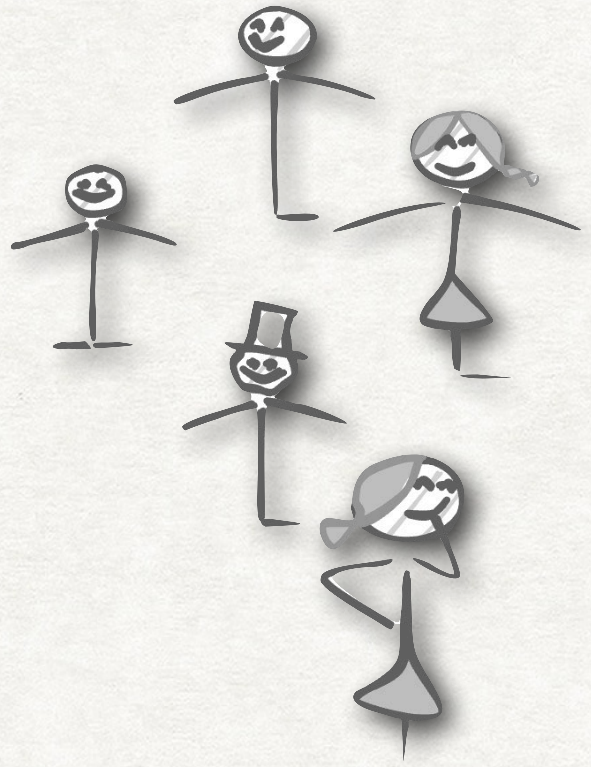
Information Retrieval

Everything, Everywhere, All at Once

Search Engines

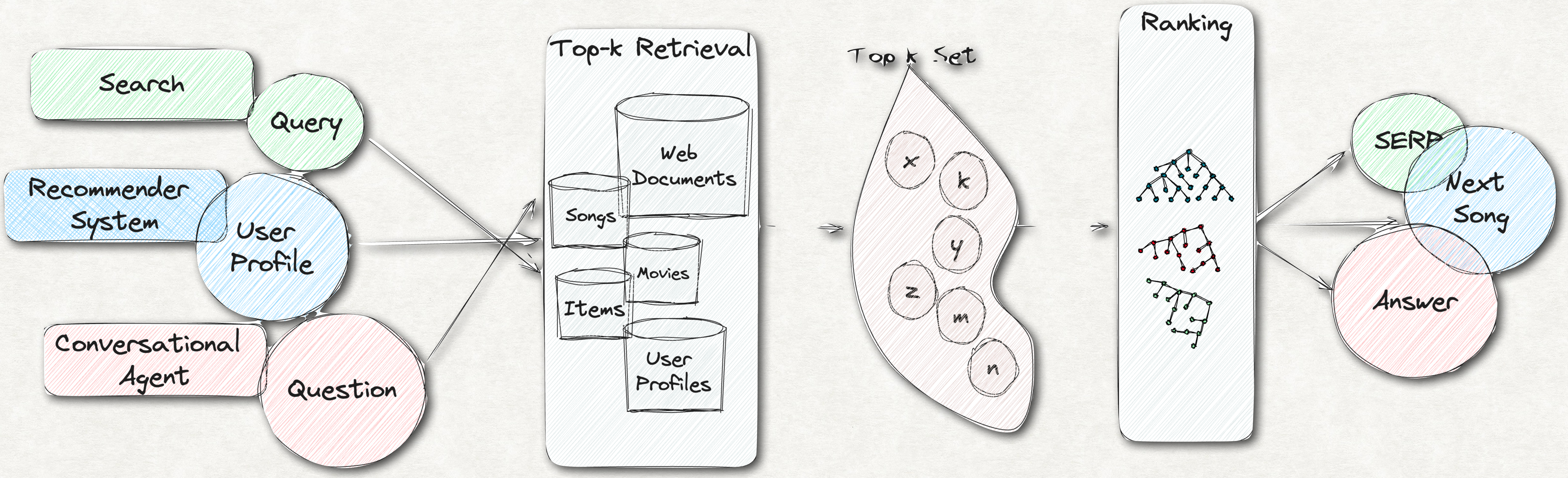
Question Answering Agents

Recommender Systems



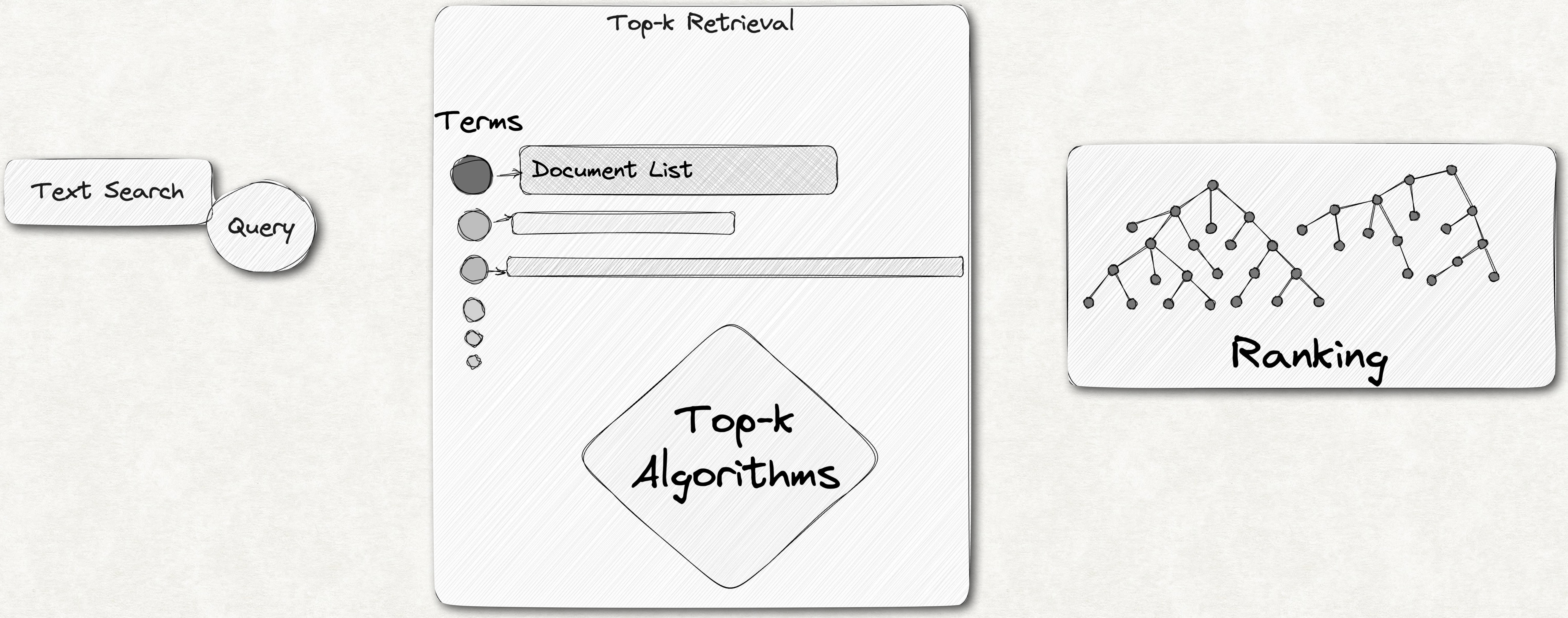
Under the Hood

Bird's Eye View



Under the Hood

Text Search

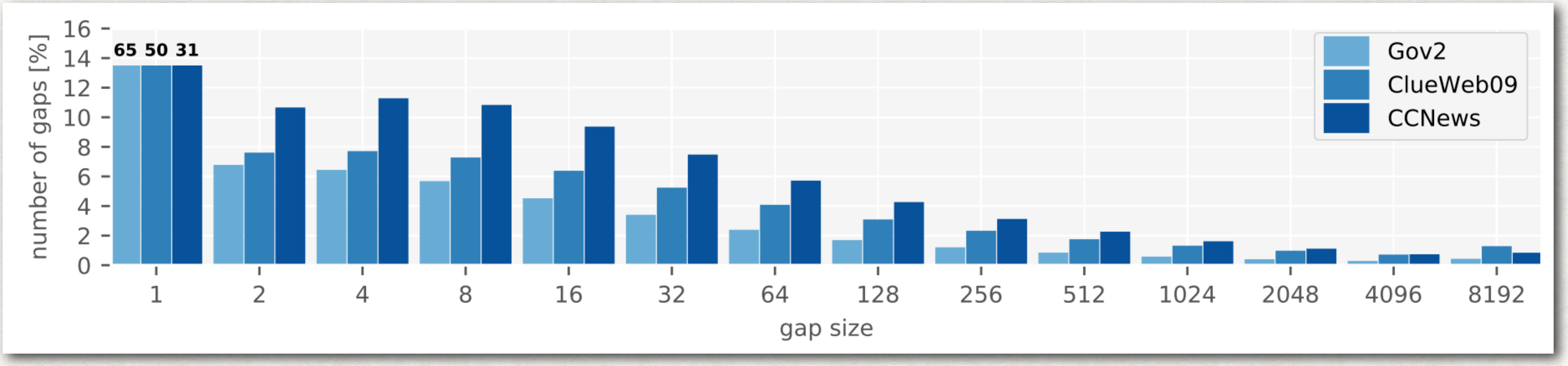


Inverted Index Compression

From Integer Codes to Inverted List and Inverted Index Compression

RQ: Compress an inverted index by optimizing storage and **decoding speed**.

- Theoretical lower-bound: $n \log_2 u/n + 1.44n$ bits for n integers of universe u.
- We can make that better for large text collections!



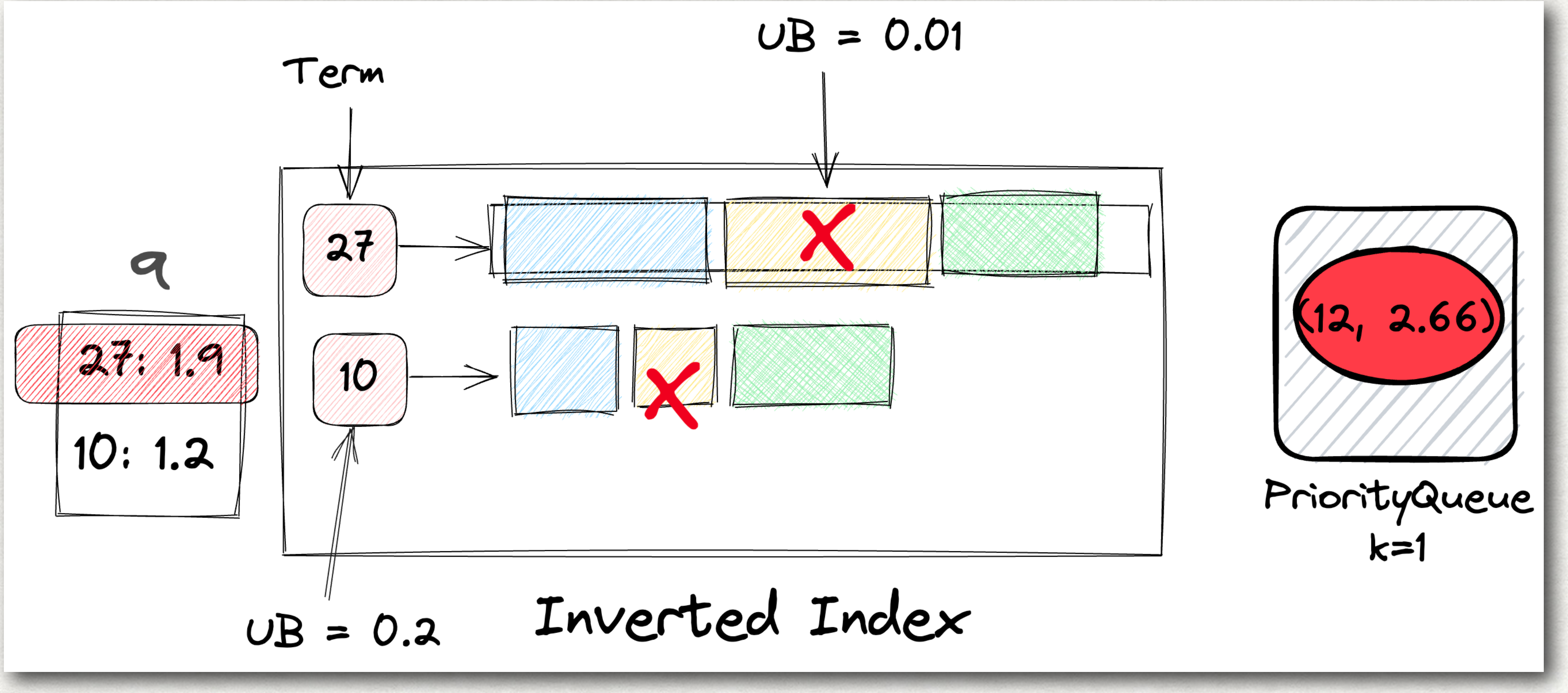
Pibiri and Venturini. "Techniques for Inverted Index Compression." ACM Computing Surveys. 2020.

Top-k Retrieval

Query Latency and Accuracy

RQ: For a query q, find the k documents from an inverted index $\mathcal{I}$ that maximize an **additive non-negative** scoring function $f(q, \cdot\,)$.

- Worst-case complexity: $\mathcal{O}(n \log k)$
- We can make that better for large text collections (Zipfian dist., non-negativity, and asymmetric query dist.)



Tonellotto, Macdonald, and Ounis. "Efficient Query Processing for Scalable Web Search." FnTIR. 2018.

GBDTs for Ranking

Compressed Data Structure and Fast Inference

RQ: Apply the function $\mathscr{T}$, a forest of n axis-aligned binary decision trees of m nodes each, to a feature vector x, by minimizing branch mispredictions

- Requires $\mathcal{O}(n \log m)$ decisions
- We can make that better for large forests!

Algorithm 2: The QuickScorer Algorithm

Input :
- $\mathbf{x}$: input feature vector
- $\mathcal{T}$: ensemble of binary decision trees, with
  - $w_0, \ldots, w_{|\mathcal{T}|-1}$: weights, one per tree
  - `thresholds`: sorted sublists of thresholds, one sublist per feature
  - `tree_ids`: tree's ids, one per threshold
  - `bitvectors`: node bitvectors, one per threshold
  - `offsets`: offsets of the blocks of triples
  - `v`: result bitvectors, one per each tree
  - `leaves`: output values, one per each tree leaf

Output:
- Final score of $\mathbf{x}$

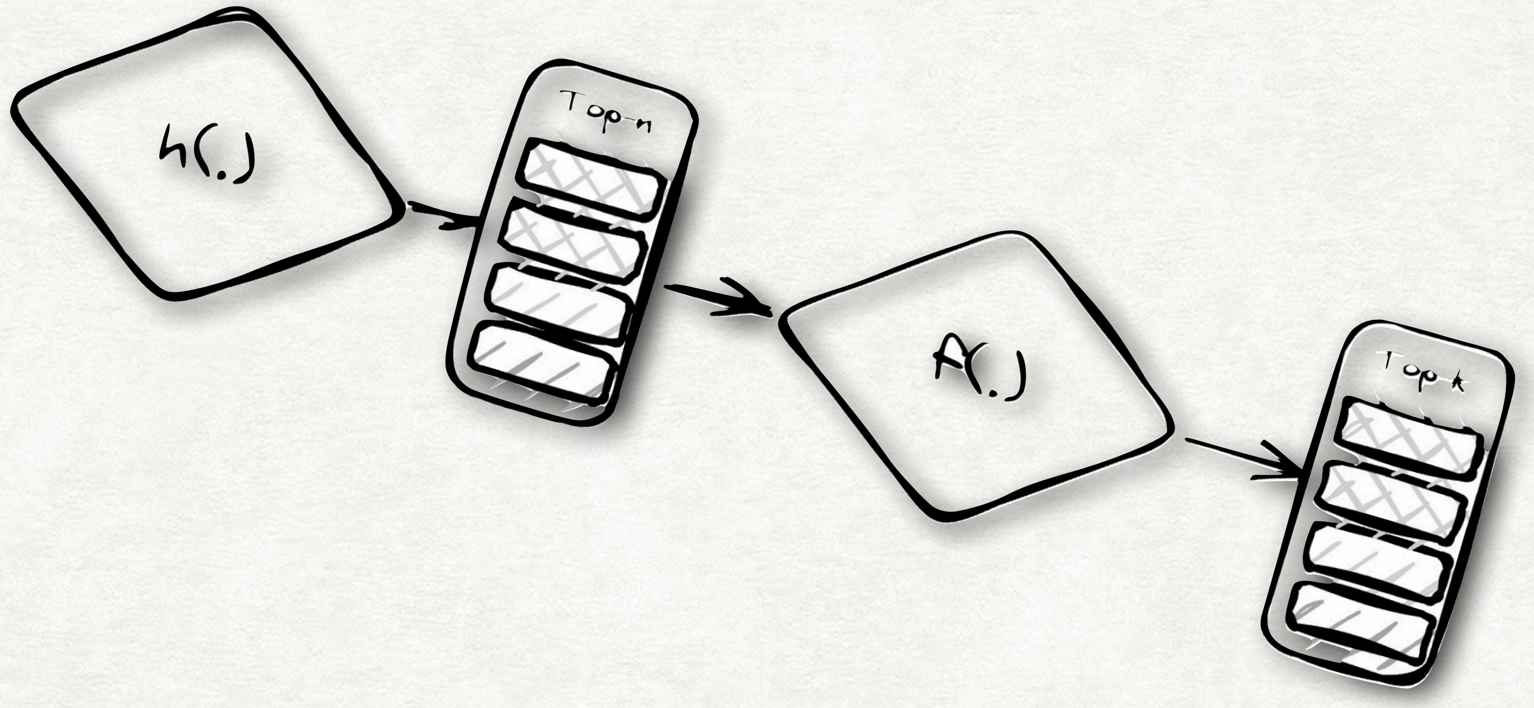
QuickScorer$(\mathbf{x},\mathcal{T})$:

foreach $h \in 0, 1, \ldots, |\mathcal{T}| - 1$ **do**

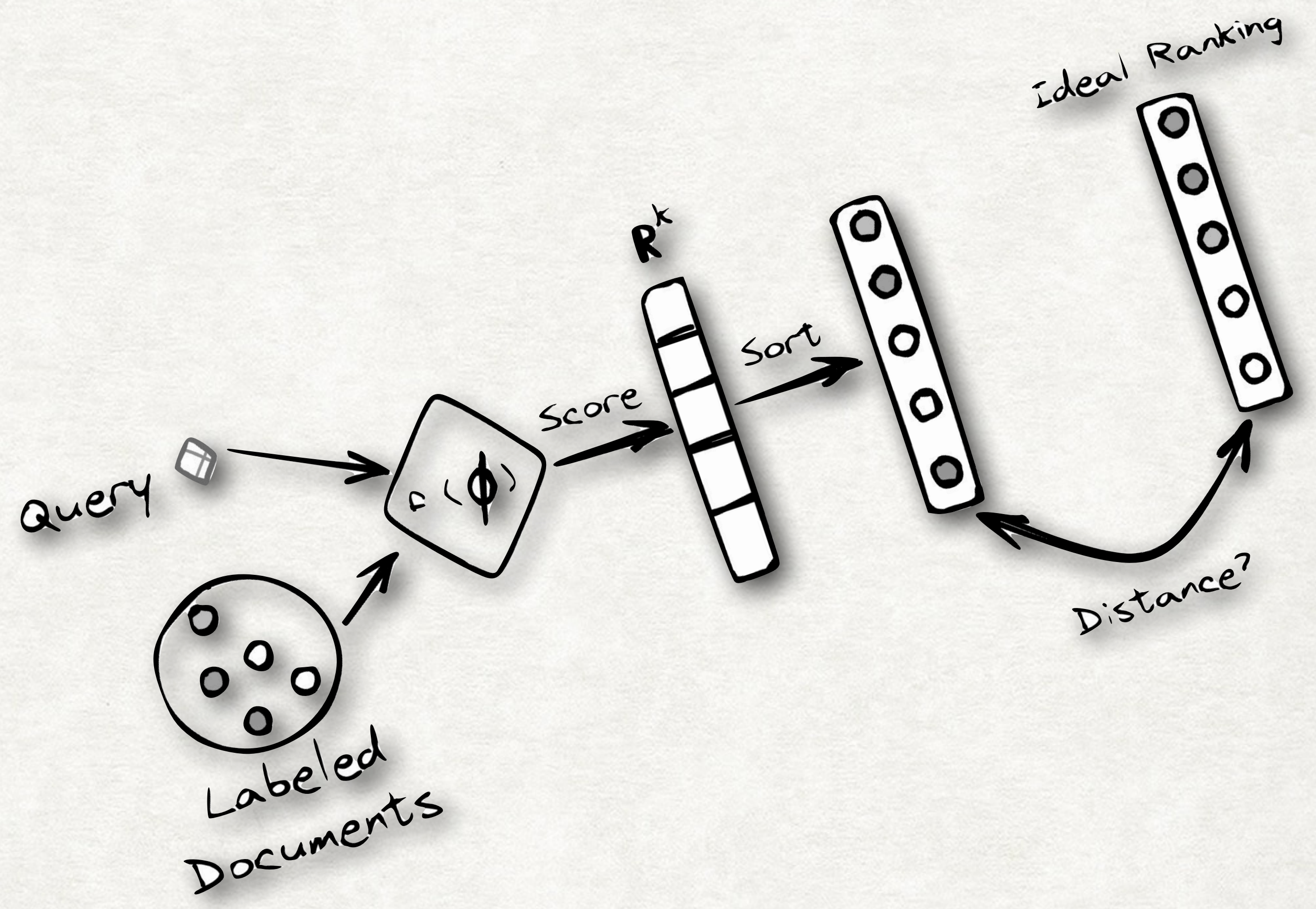
Bruch, Lucchese, and Nardini. "Efficient and Effective Tree-based and Neural Learning to Rank." FnTIR. 2023

Other Notable Lines of Research

Wait! There is more!



- Multi-stage Ranking Systems
  - Zamani et al. "*Stochastic Retrieval-Conditioned Reranking.*" ICTIR. 2022.
- Learning Ranking Functions
  - Bengs et al. "*Preference-based Online Learning with Dueling Bandits: A Survey.*" JMLR. 2021.
- Evaluation Measures and Statistical Tests
  - Ferrante, Ferro, and Pontarollo. "A General Theory of IR Evaluation Measures." TKDE. 2019.



“

PISA is capable of returning the **top 10** documents with an average latency in the range of **10-40 milliseconds** on a collection containing **50 million web documents**.

”

Mallia, Siedlaczek, and Suel. *“An Experimental Study of Index Compression and DAAT Query Processing Methods.” ECIR 2019*

Observation I

Plenty of problems ranging from data structures, Information theory, algorithms, learning theory, Theory, and systems.

Observation II

Formalizing problems leads to principled, **robust** solutions

Act II: Magic

Neural Networks, Everything Else is a Distraction

Enter Neural Networks

A new era in Text Ranking

A Deep Look into Neural Ranking Models for Information Retrieval

Jiafeng Guo[a,b], Yixing Fan[a,b], Liang Pang[a,b], Liu Yang[c], Qingyao Ai[c], Hamed Zamani[c], Chen Wu[a,b], W. Bruce Croft[c], Xueqi Cheng[a,b]

[a] University of Chin
[b] CAS Key Lab of Network D
Technology, Chine
[c] Center for Intelligent Informa

Pretrained Transformers for Text Ranking: BERT and Beyond

Jimmy Lin,[1] Rodrigo Nogueira,[1] and Andrew Yates[2,3]

[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] University of Amsterdam
for Informatics

gust 20, 2021

s.IR] 25 Jan 2023

Conversational Information Seeking

An Introduction to Conversational Search, Recommendation, and Question Answering

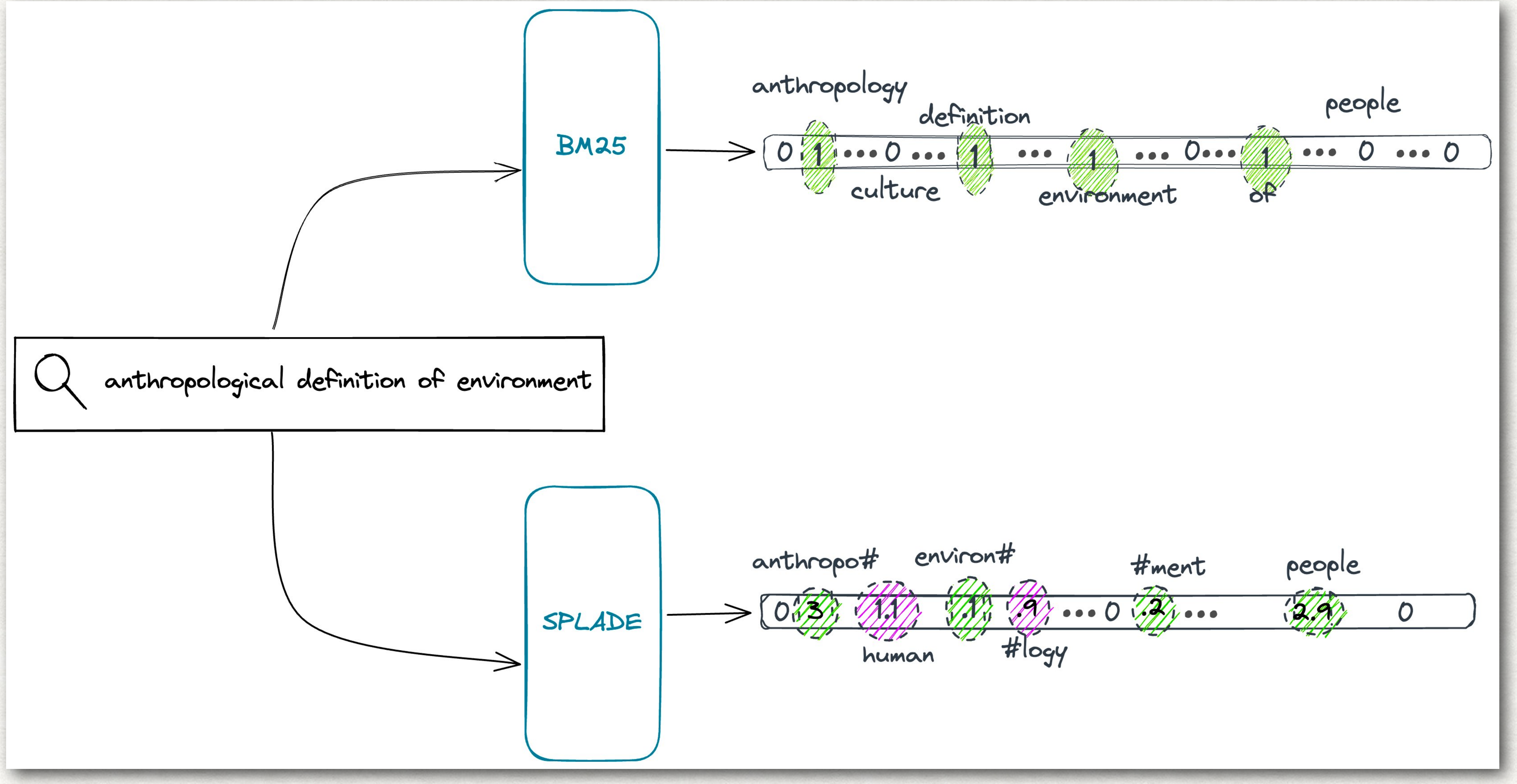
Suggested Citation: Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski (2023), "Conversational Information Seeking", : Vol. xx, No. xx, pp 1–222. DOI: 10.1561/XXXXXXXXX.

Mean Reciprocal Rank on the MS MARCO v1 (Passage) dataset

	Test MRR@10
BM25	0.218
IRNet (reranking)	0.281
BM25 (retrieval) and BERT (reranking)	0.365
SOTA (2023-09)	0.450

Bag of Learnt Words

Learning term importance

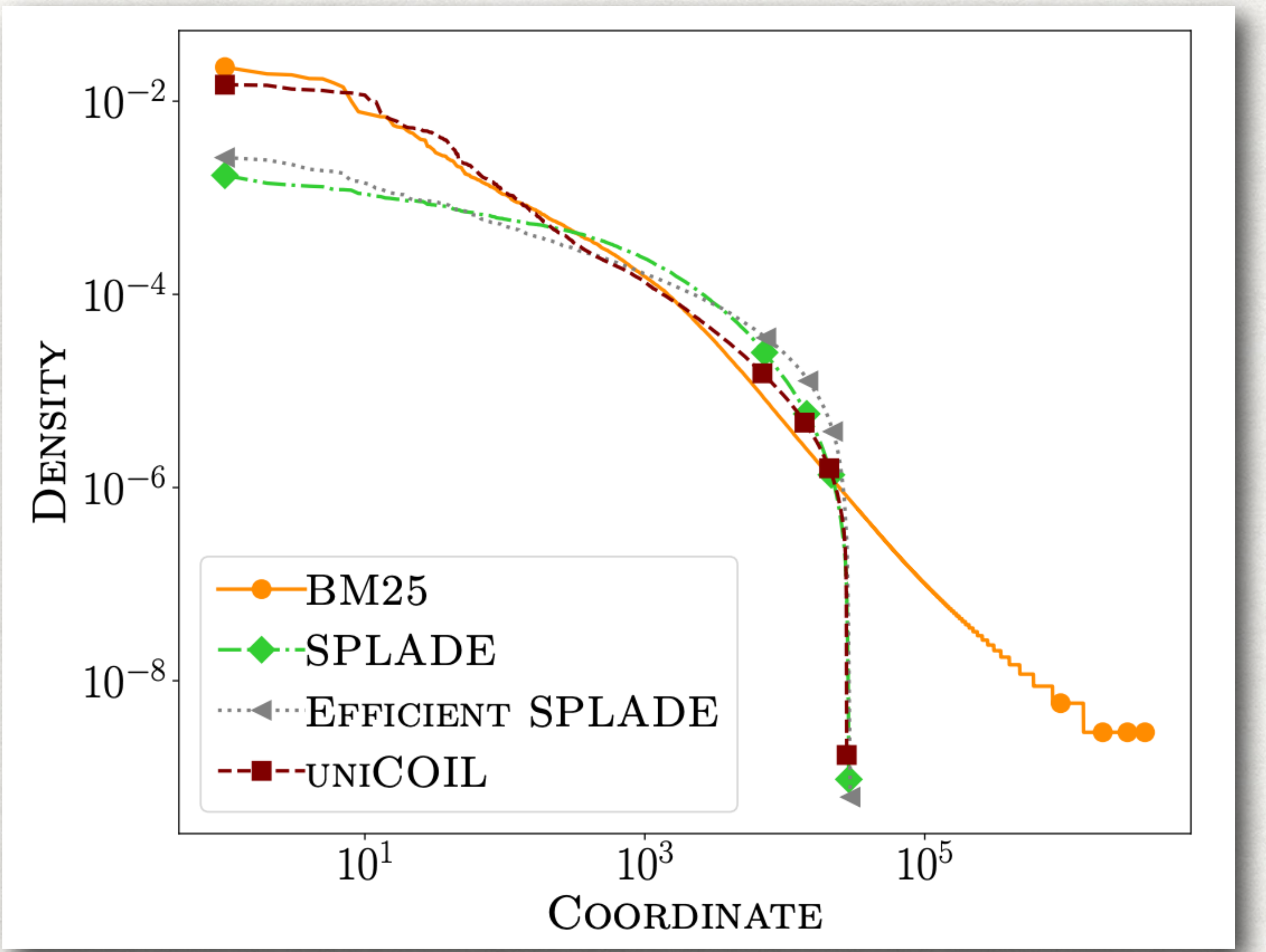


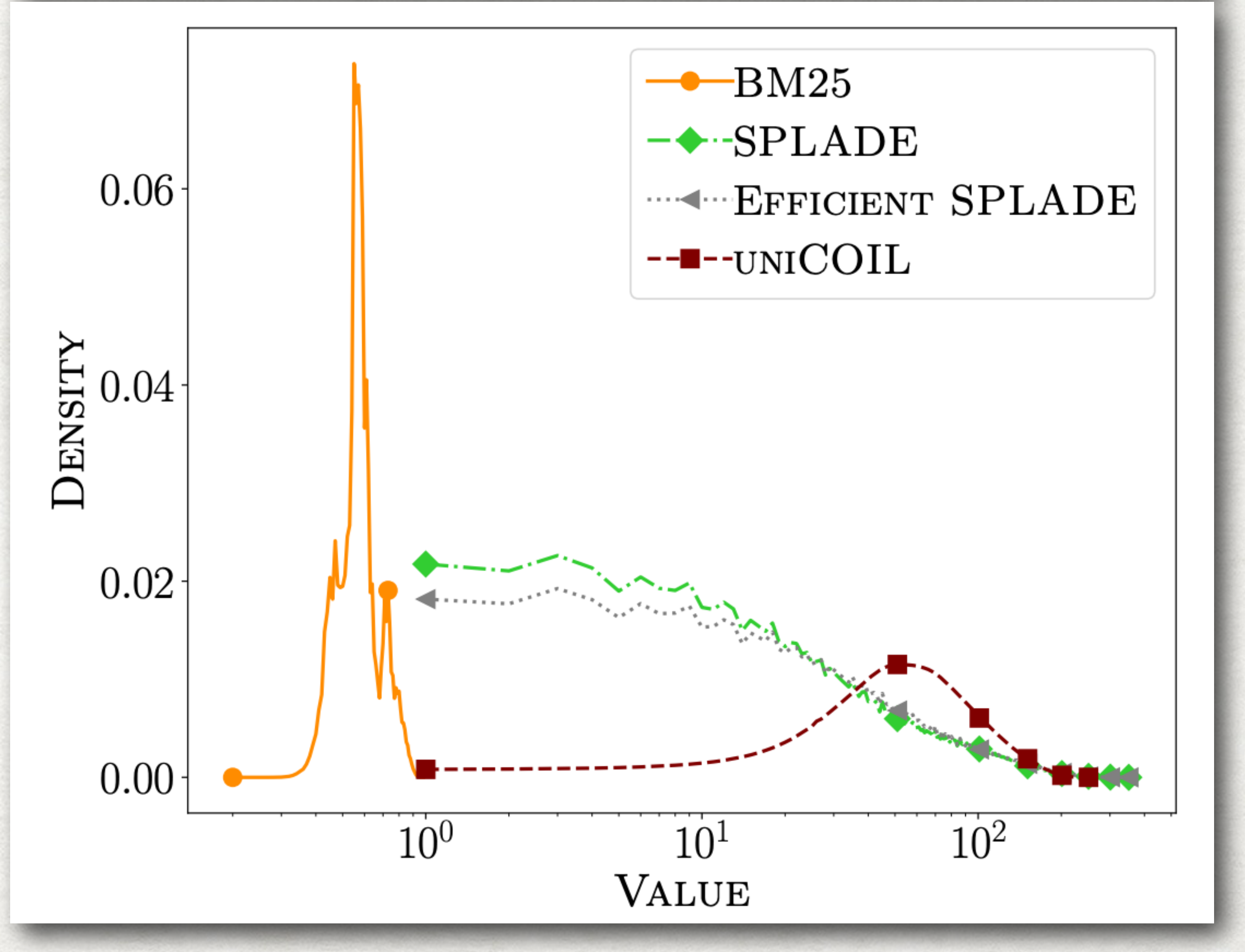
Mean Reciprocal Rank on the MS MARCO v1 (Passage) dataset

	Test MRR@10
BM25	0.218
SPLADE	0.383
SOTA (2023-09)	0.450

Formal et al. "*SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval.*"

Distribution of non-zero coordinates





Distribution of values

What is wrong with that picture?

- **Limitations to efficiency**
  - Inverted lists violate assumptions underlying compression, dynamic pruning algorithms
- **Limitations to effectiveness**
  - Queries and documents *must* have different distributions
  - Vectors *must* be non-negative and discretized

MS MARCO v1 (Passage) dataset

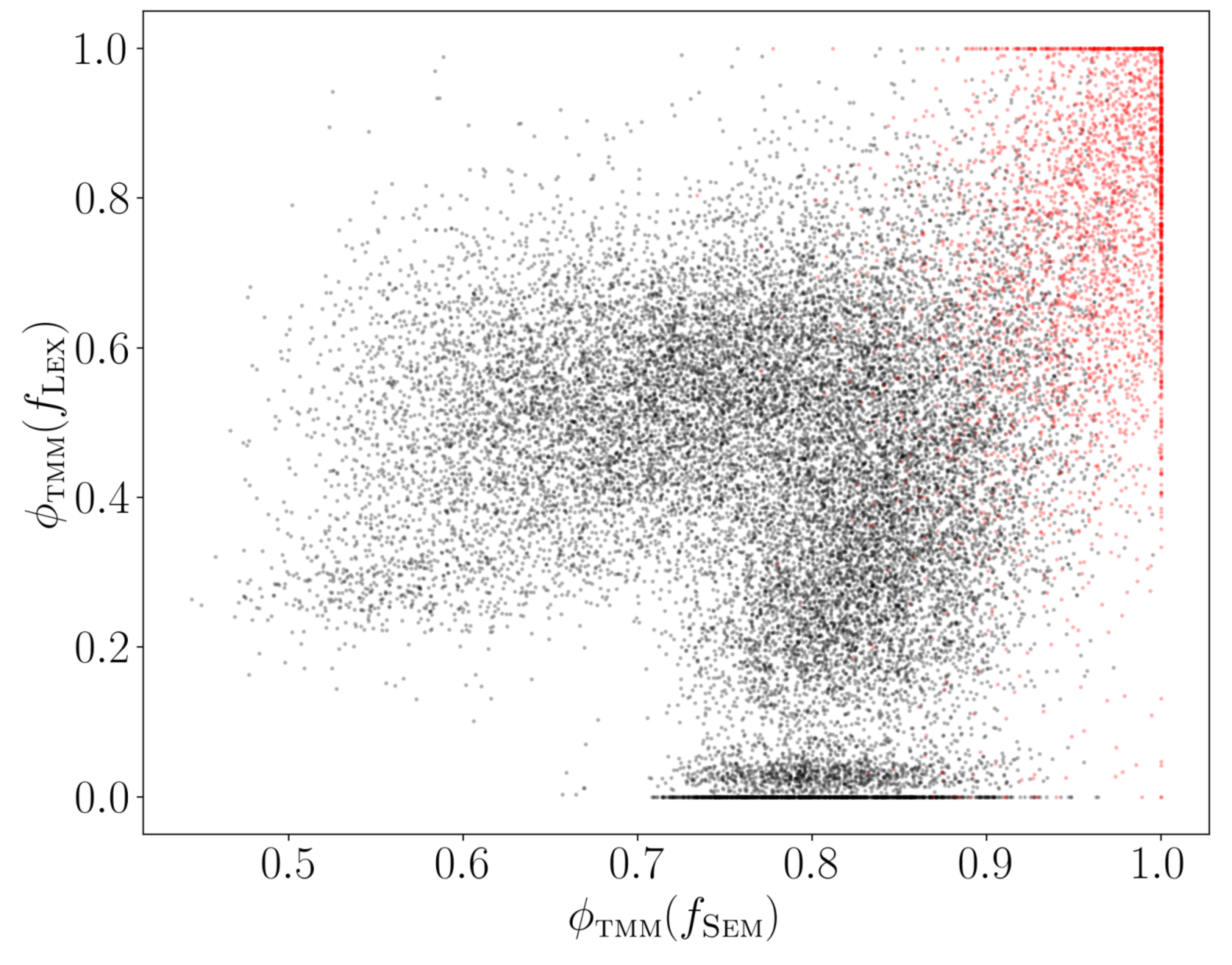
	WAND Query Latency (ms)
BM25	~35
SPLADE	~1000

Bruch et al. "*An Approximate Algorithm for Maximum Inner Product Search over Streaming Sparse Vectors.*" ACM TOIS. 2023.

Fusion of Vectors

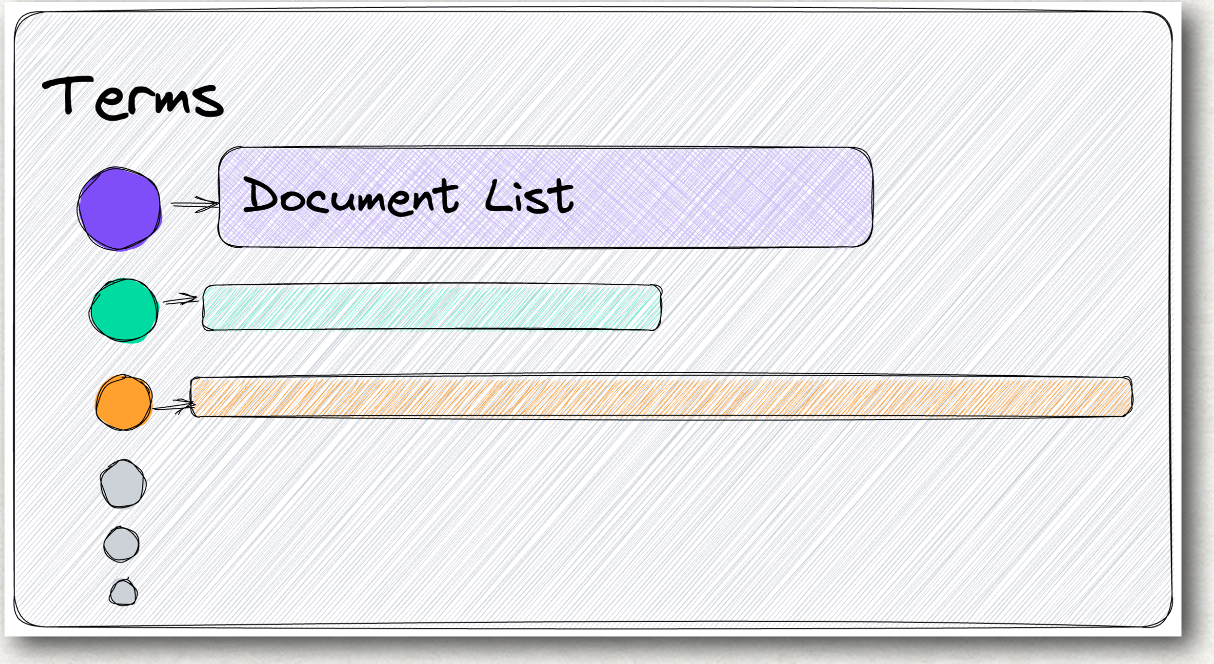
Lexical-Semantic and Multimodal Search

- Lexical and semantic models encode different information about text!
- Multimodal data need retrieval over joint representations!

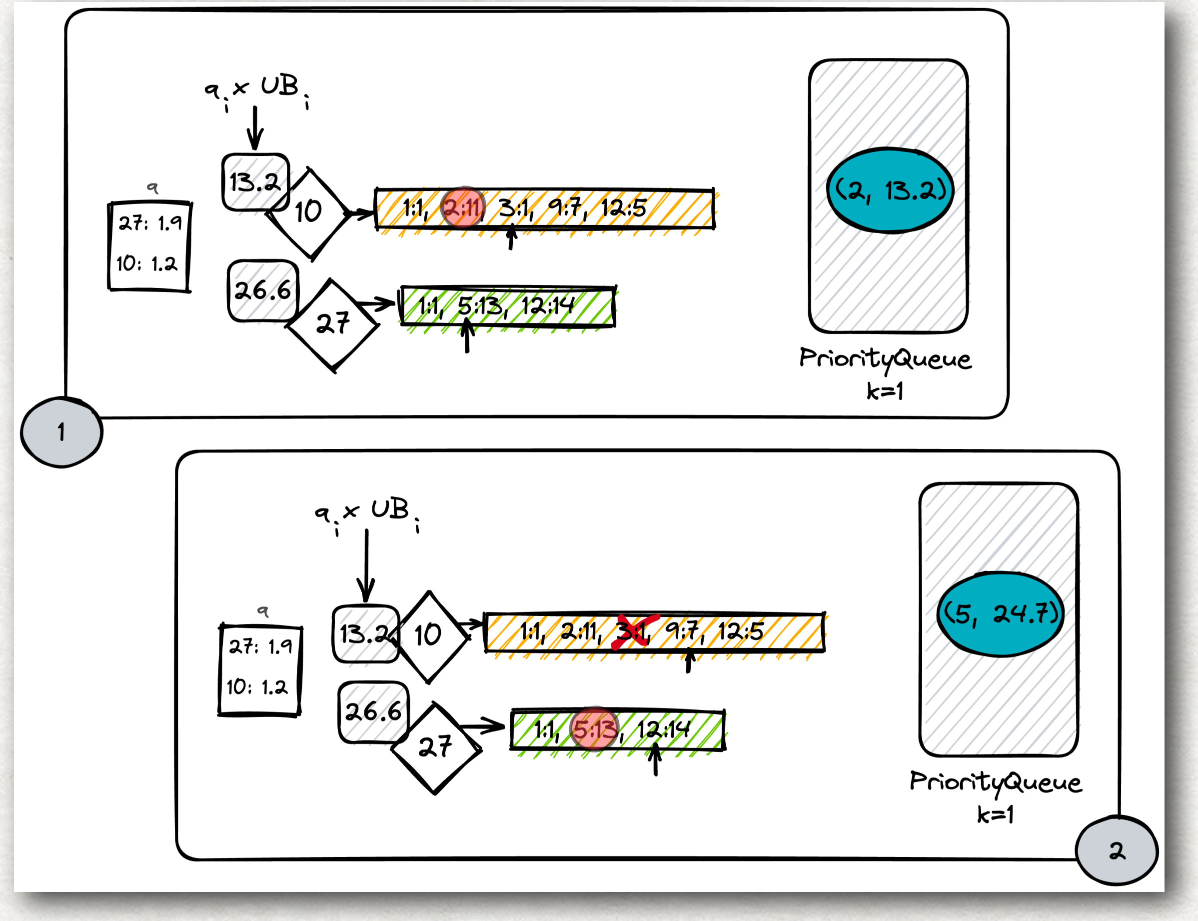


Bruch et al. "An Analysis of Fusion Functions for Hybrid Retrieval." ACM TOIS. 2023.

Solutions Abound

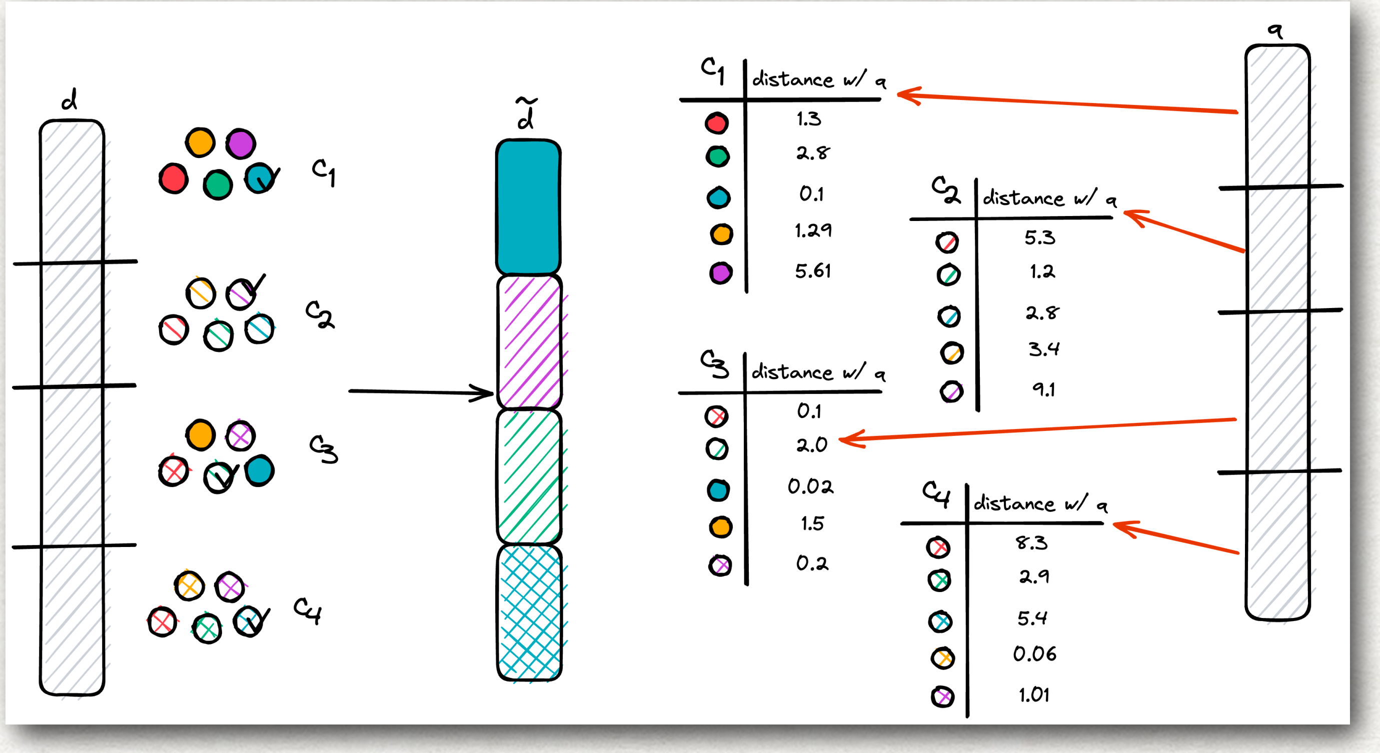


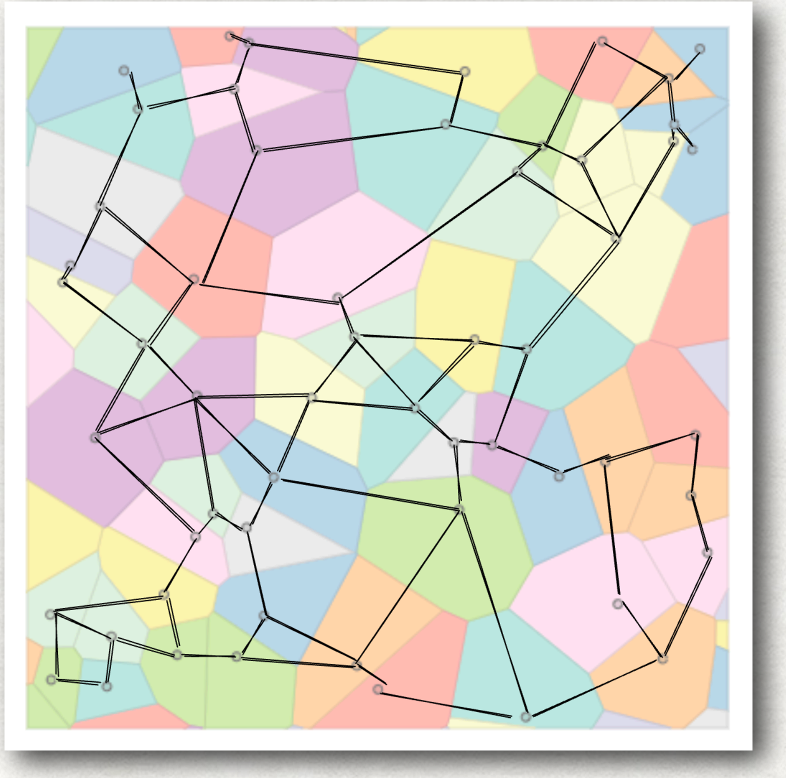
Inverted Index



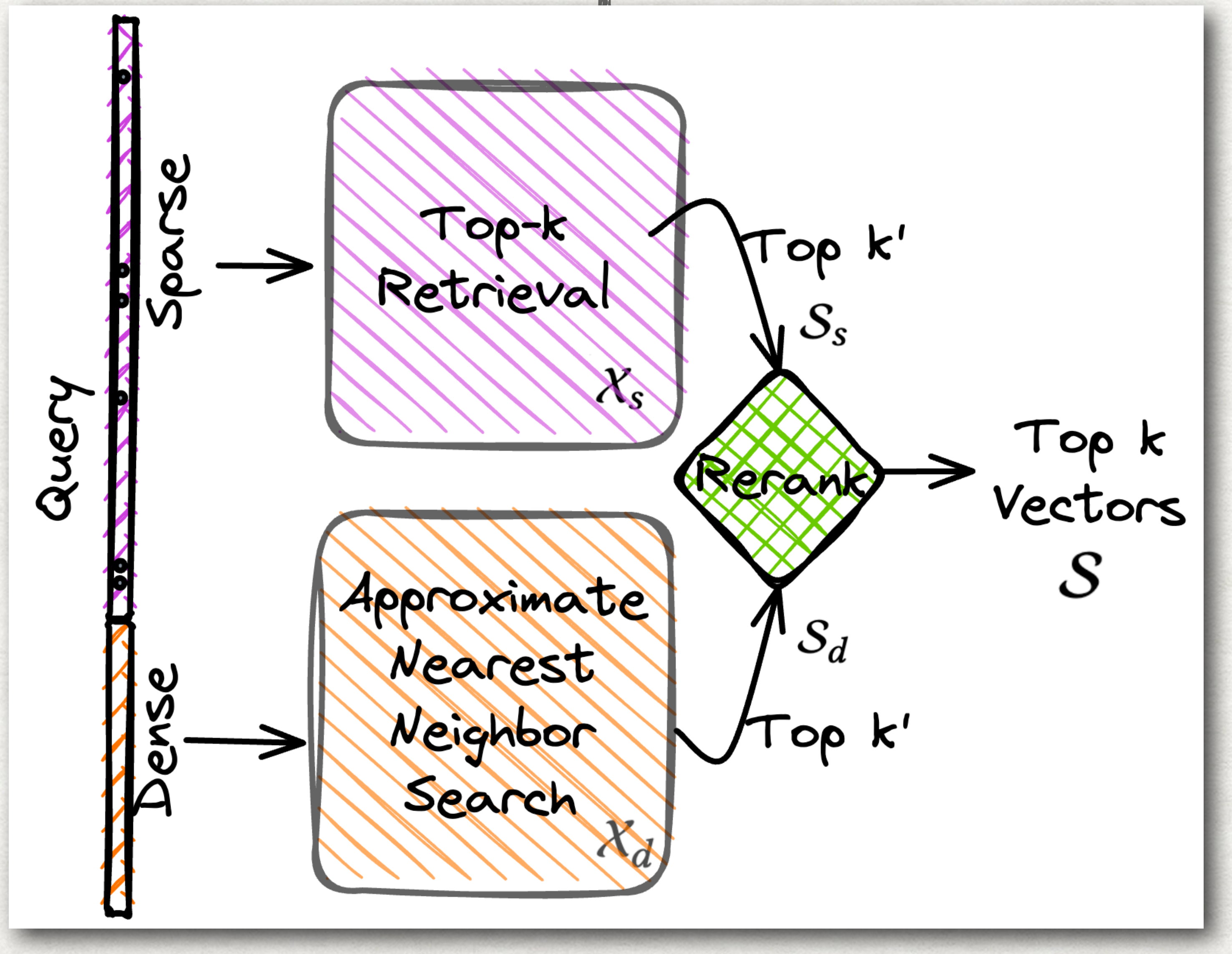
Top-k Retrieval

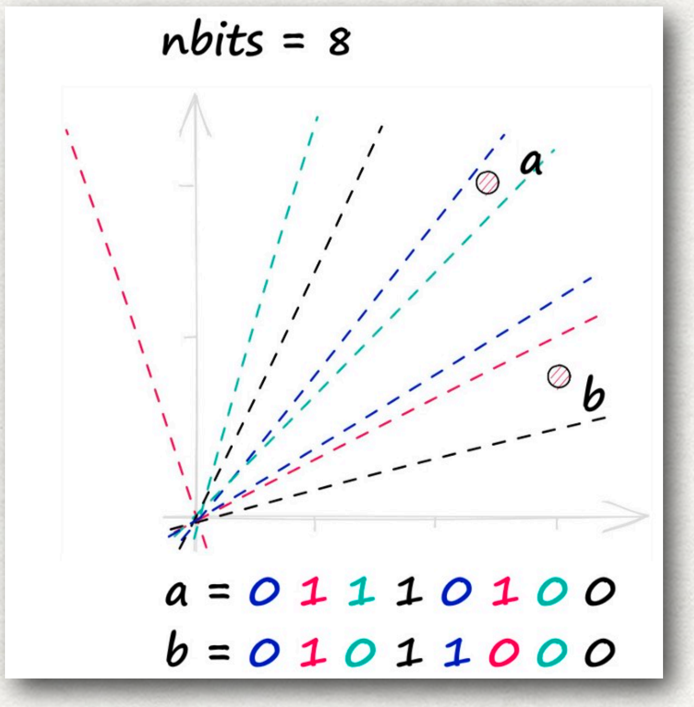
Product Quantization





Graph-based ANN

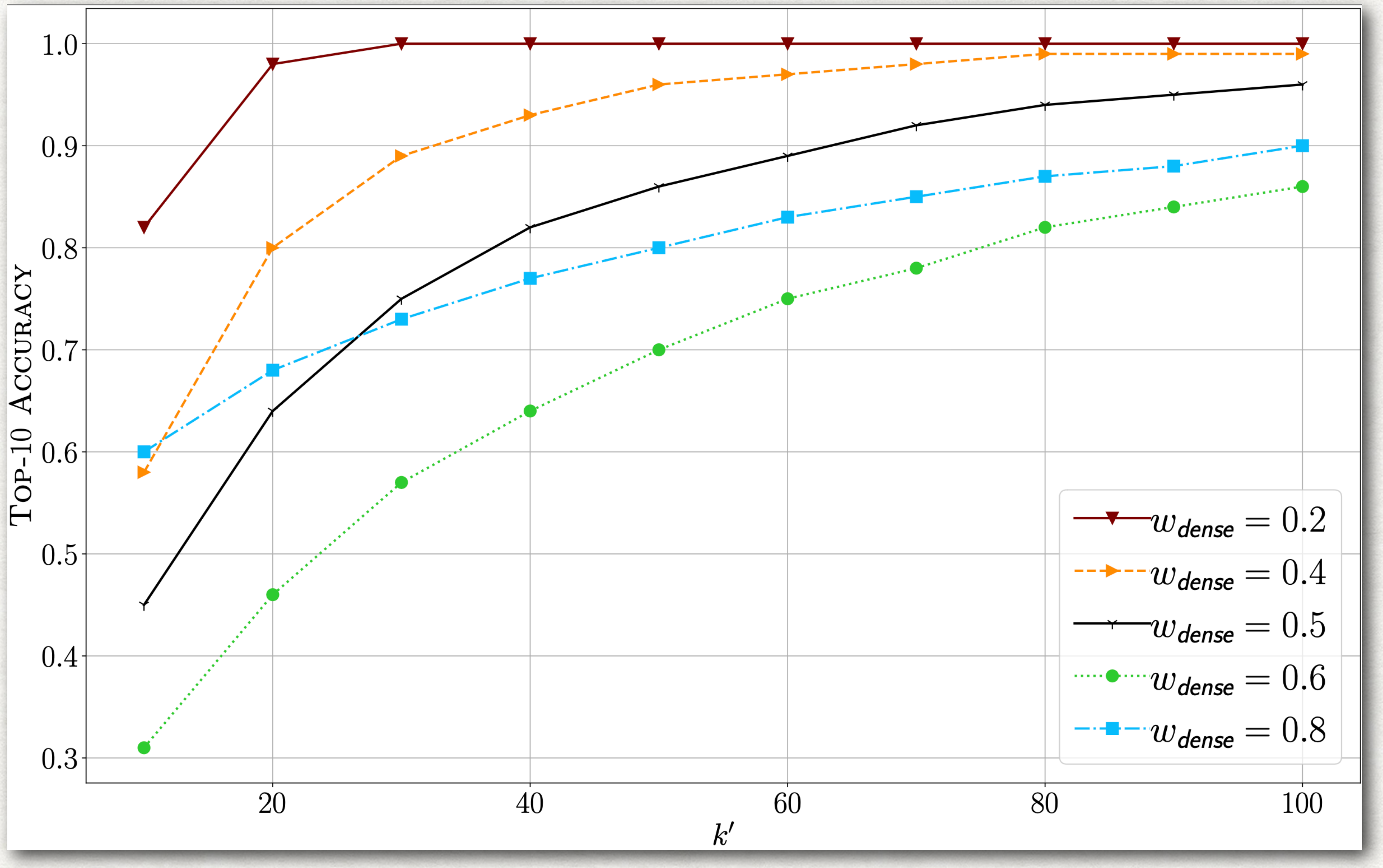




Random Projections

What is wrong with that picture?

- **Limitations to efficiency**
  - We *must* retrieve $k' \gg k$ to compensate for the separation of systems
- **Limitations to effectiveness**
  - Poor retrieval quality when vectors are uncorrelated

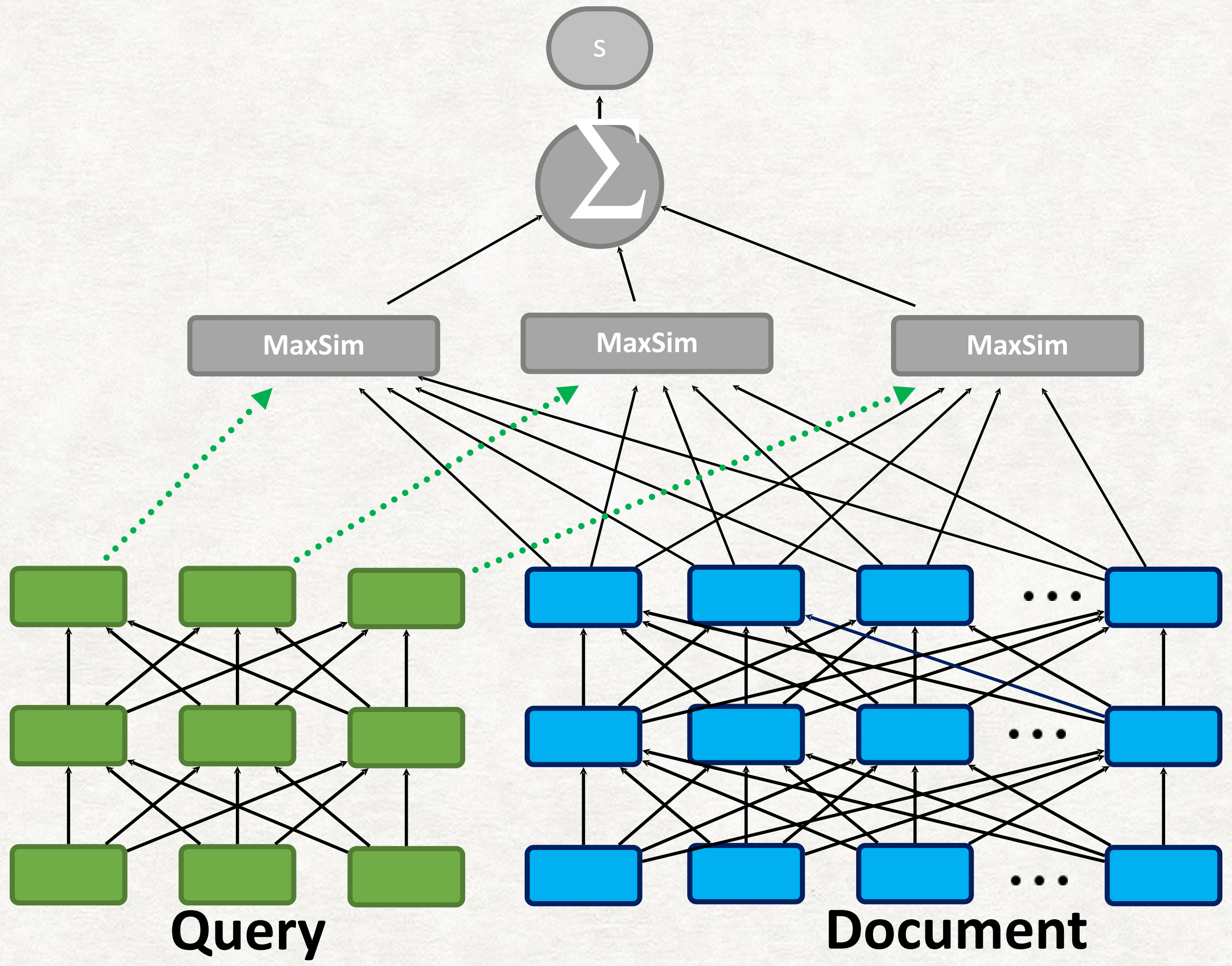


Dense vectors in $\mathbb{R}^{64}$ drawn from the exponential distribution and sparse vectors from $\mathbb{R}^{1000}$ with average of 16 non-zero coordinates.

~~Bruch et al. *"Bridging Dense and Sparse Maximum Inner Product Search."* Under Review.~~

Representing Documents as a Matrix

Bag of Vectors



$$\underset{X}{argmax} \|QX\|_{\infty}$$

Mean Reciprocal Rank on the MS MARCO v1 (Passage) dataset

	Test MRR@10
BM25	0.218
SPLADE	0.383
COLBERTV2	0.397
SOTA (2023-09)	0.450

Santhanam et al. "*ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction.*" NAACL. 2022

What is wrong with that picture?

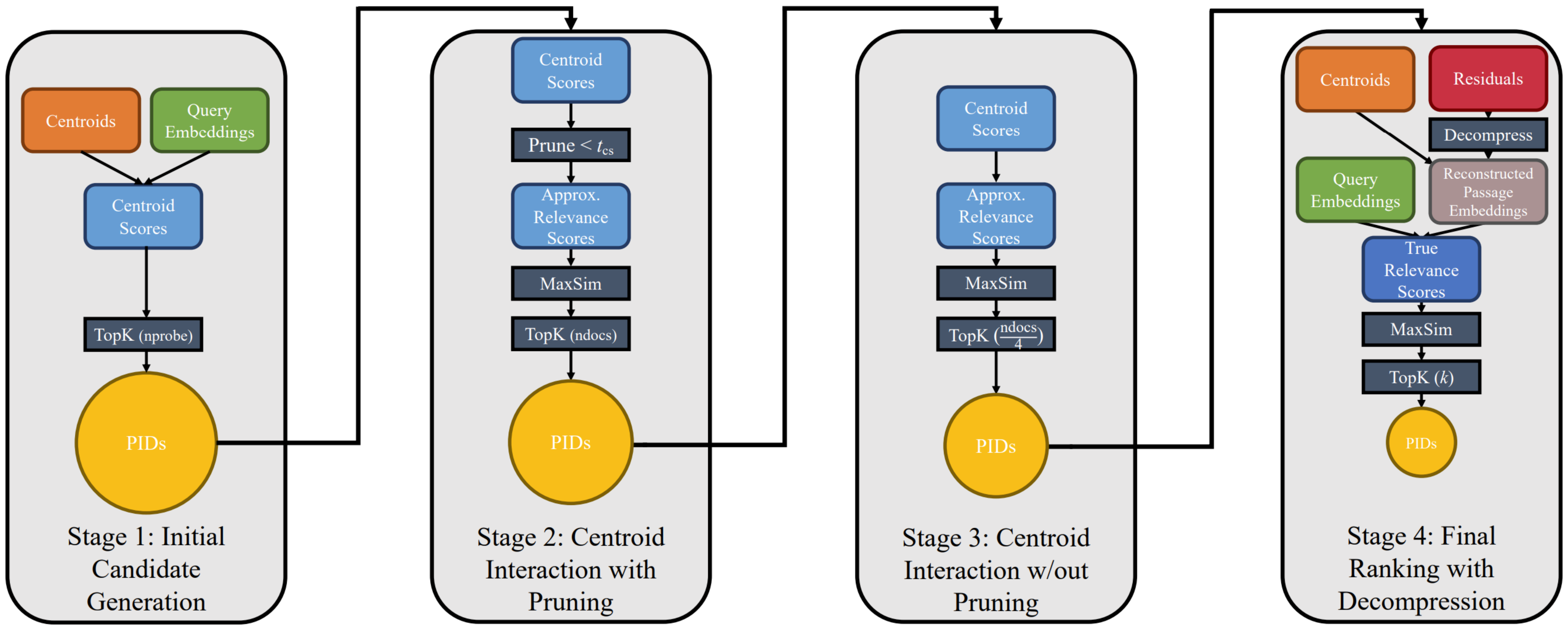


Figure 5: The PLAID scoring pipeline. The first stage generates an initial set of candidate passages using the centroids. Next the second and third stages leverage centroid pruning and centroid interaction respectively to refine the candidate set. Then the last stage performs full residual decompression to obtain the final passage ranking. We use the hyperparameter ndocs to specify the number of candidates returned by Stage 2, and in our experiments we have Stage 3 output $\frac{\text{ndocs}}{4}$ passages.

Santhanam et al. "*PLAID: An Efficient Engine for Late Interaction Retrieval.*" CIKM. 2022

Observation I

Existing algorithmic tools enable discovery of promising ideas, but they **shape your view** and future research

Observation II

Heuristics are temporary, **fragile** solutions

Act III: Maximum Inner Product Search

Example of a Hard Problem

Everything is a Vector is Everything

Multi-modality is Single-modality

Ranking is Retrieval

Retrieval is

$$\underset{v \in \mathcal{D}}{\overset{(k)}{argmax}}\, q^T v$$

Subproblems

Compression

Indexing

retrieval

Euclidean Distance

Subproblems

Compression

Indexing

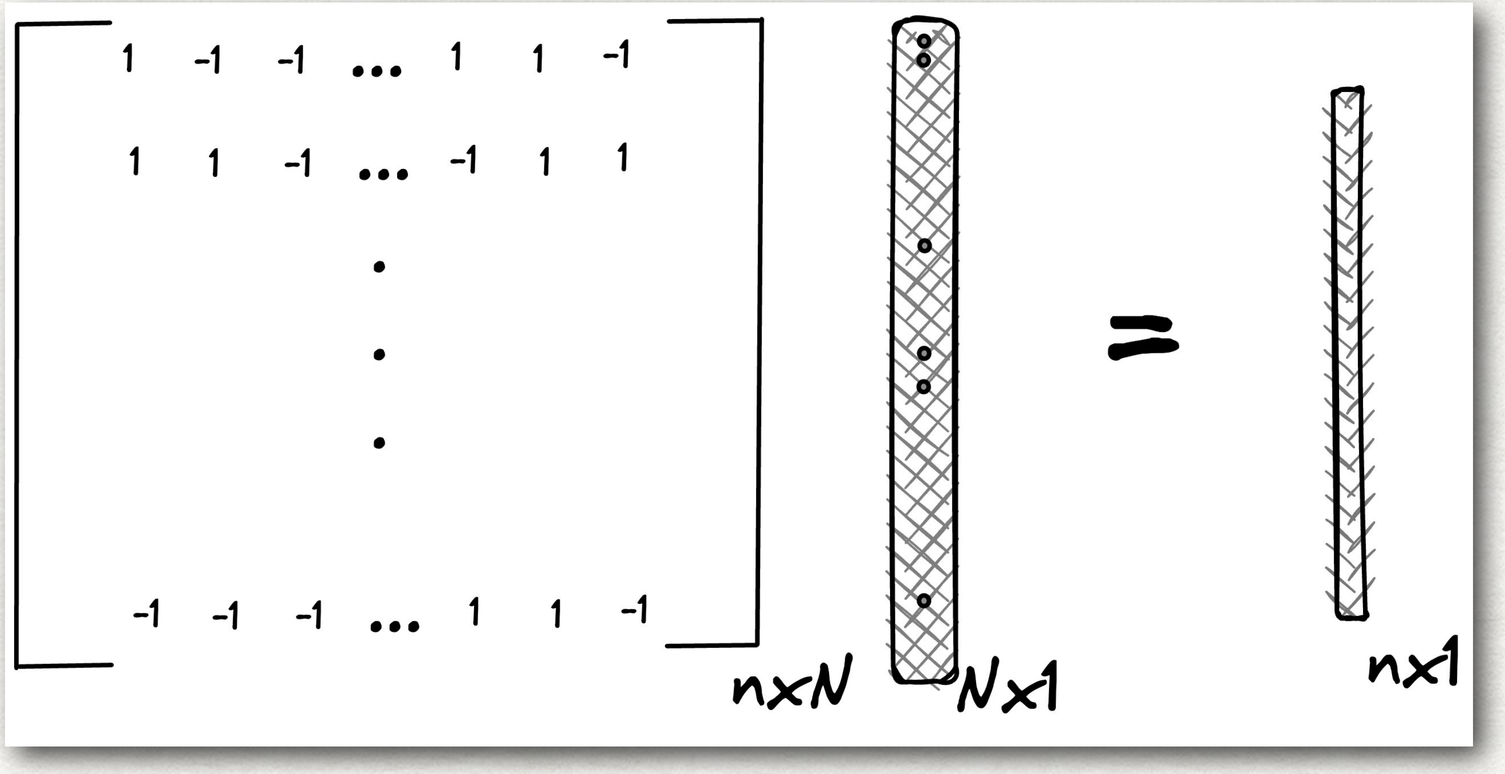
retrieval

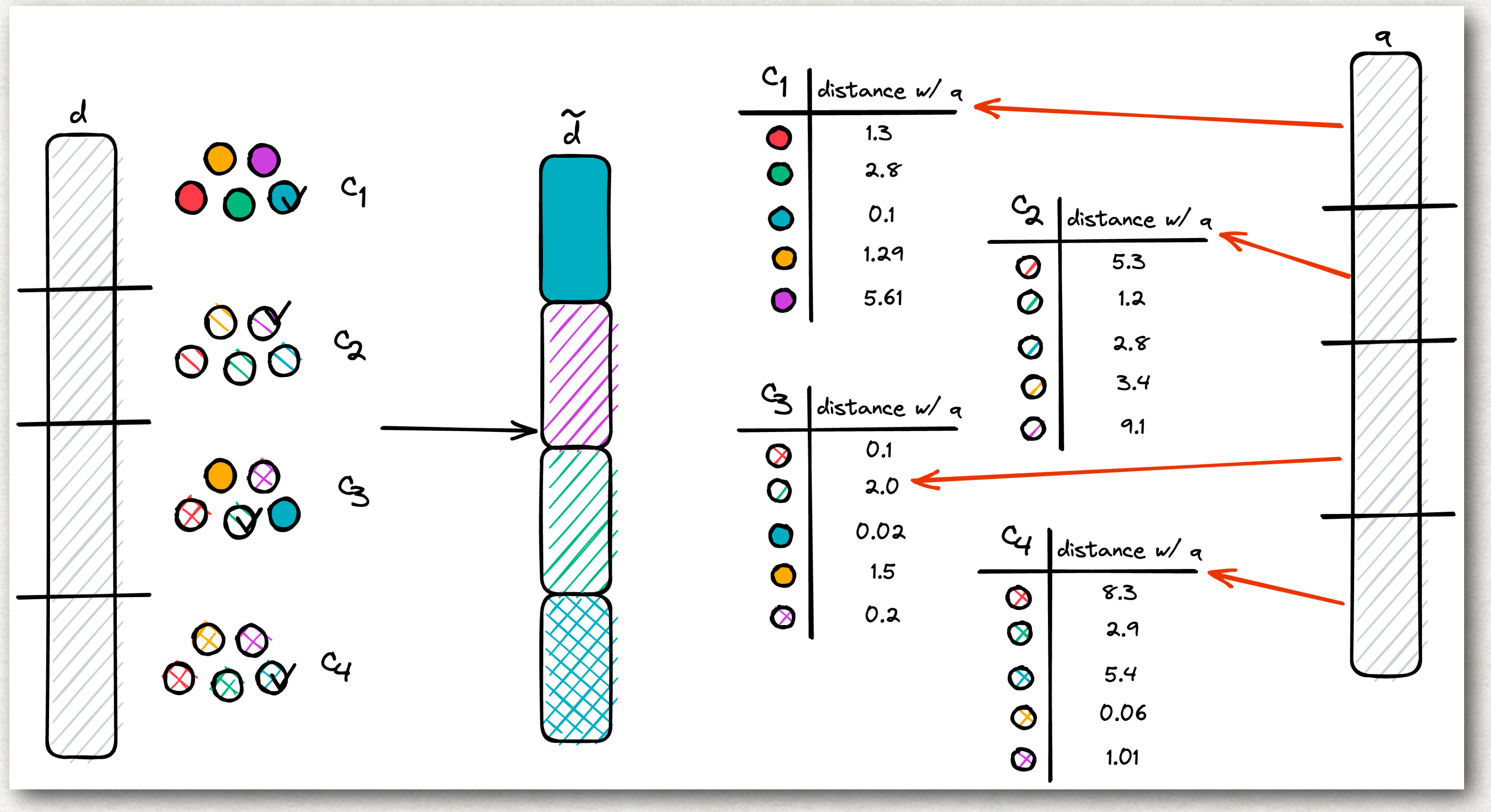
Vector Compression

- **RQ**: Find a transformation $f : \mathbb{R}^N \rightarrow \mathbb{R}^n$ that preserves Euclidean distance between vectors:

$$\|f(x) - f(y)\|_2 \approx \|x - y\|_2$$

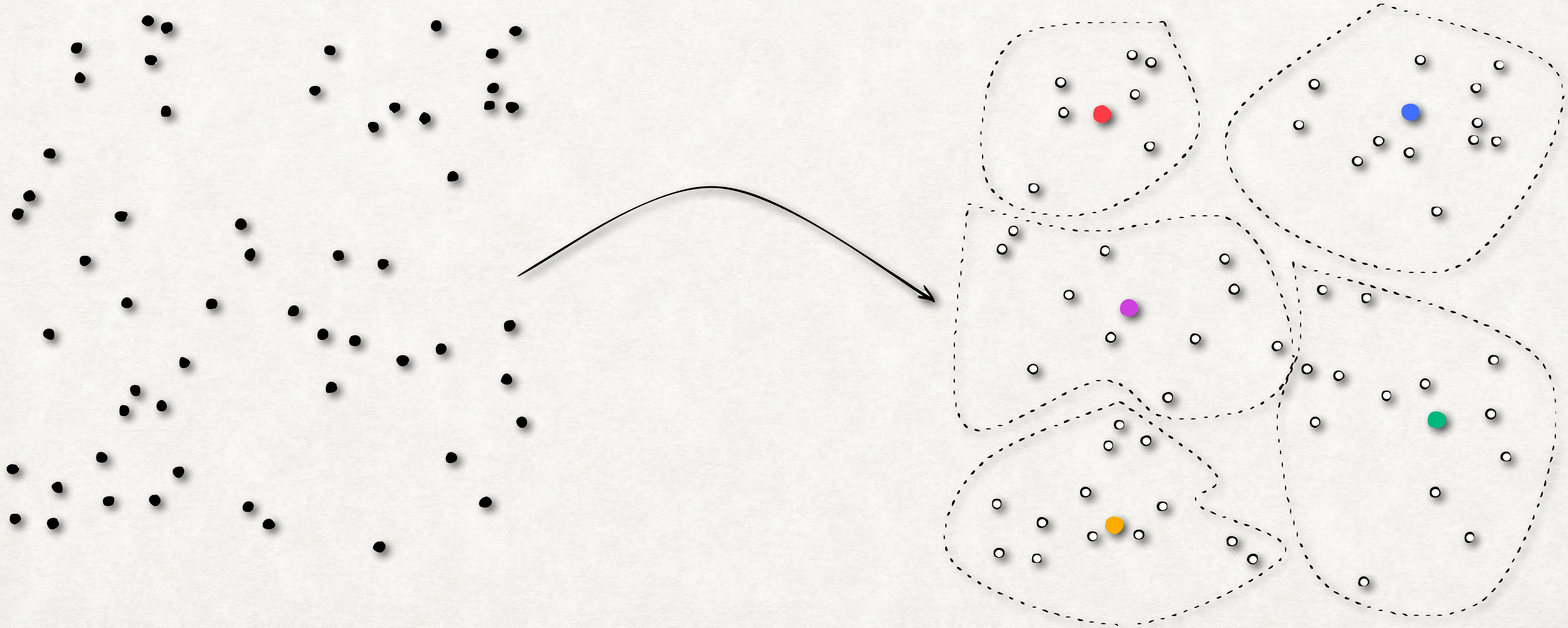
Linear Sketches such as JL Transform





Product Quantization

Indexing using Space Partitioning

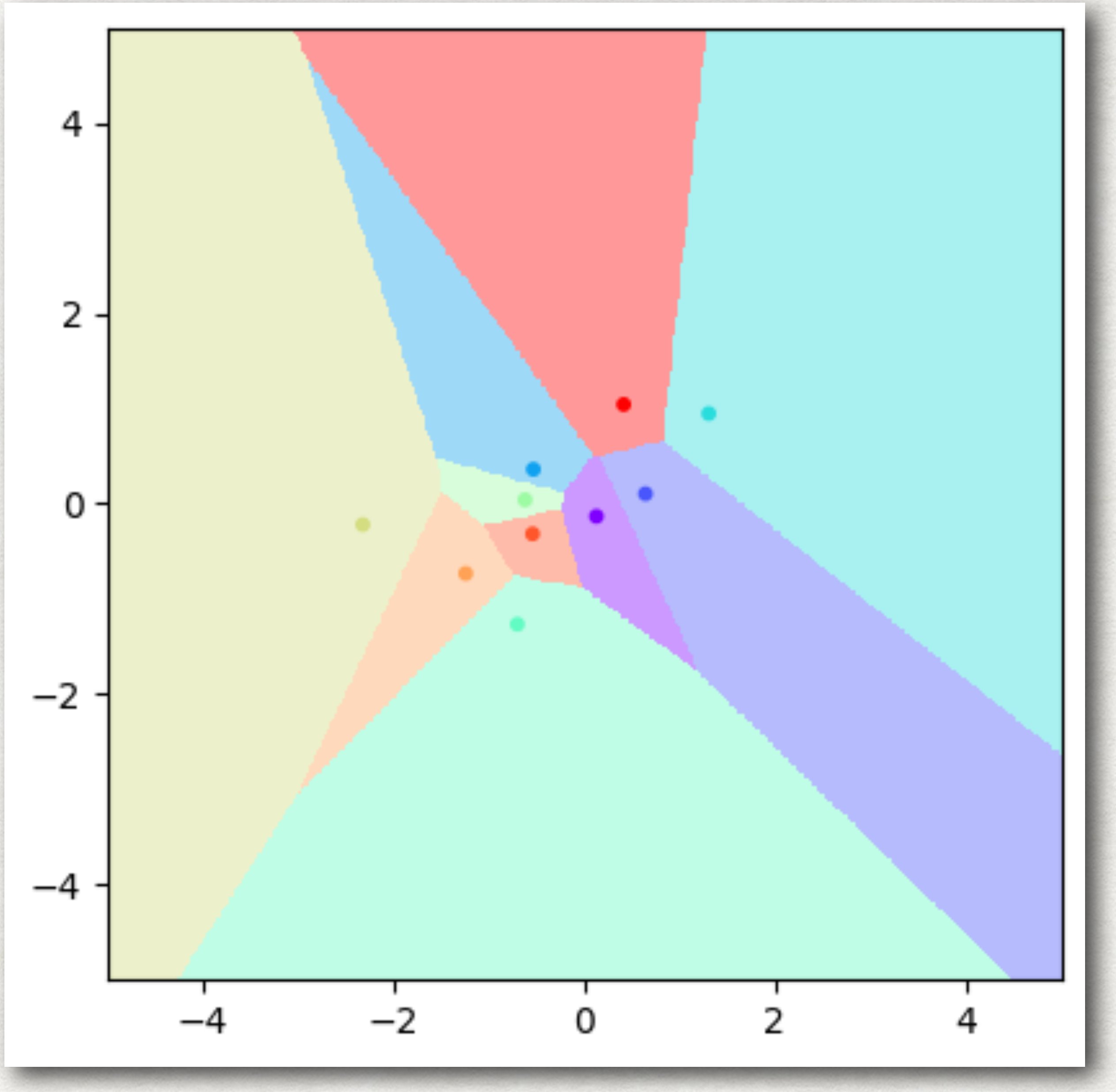


Partitioning

- **RQ**: Find partitions that approximate Voronoi cells:

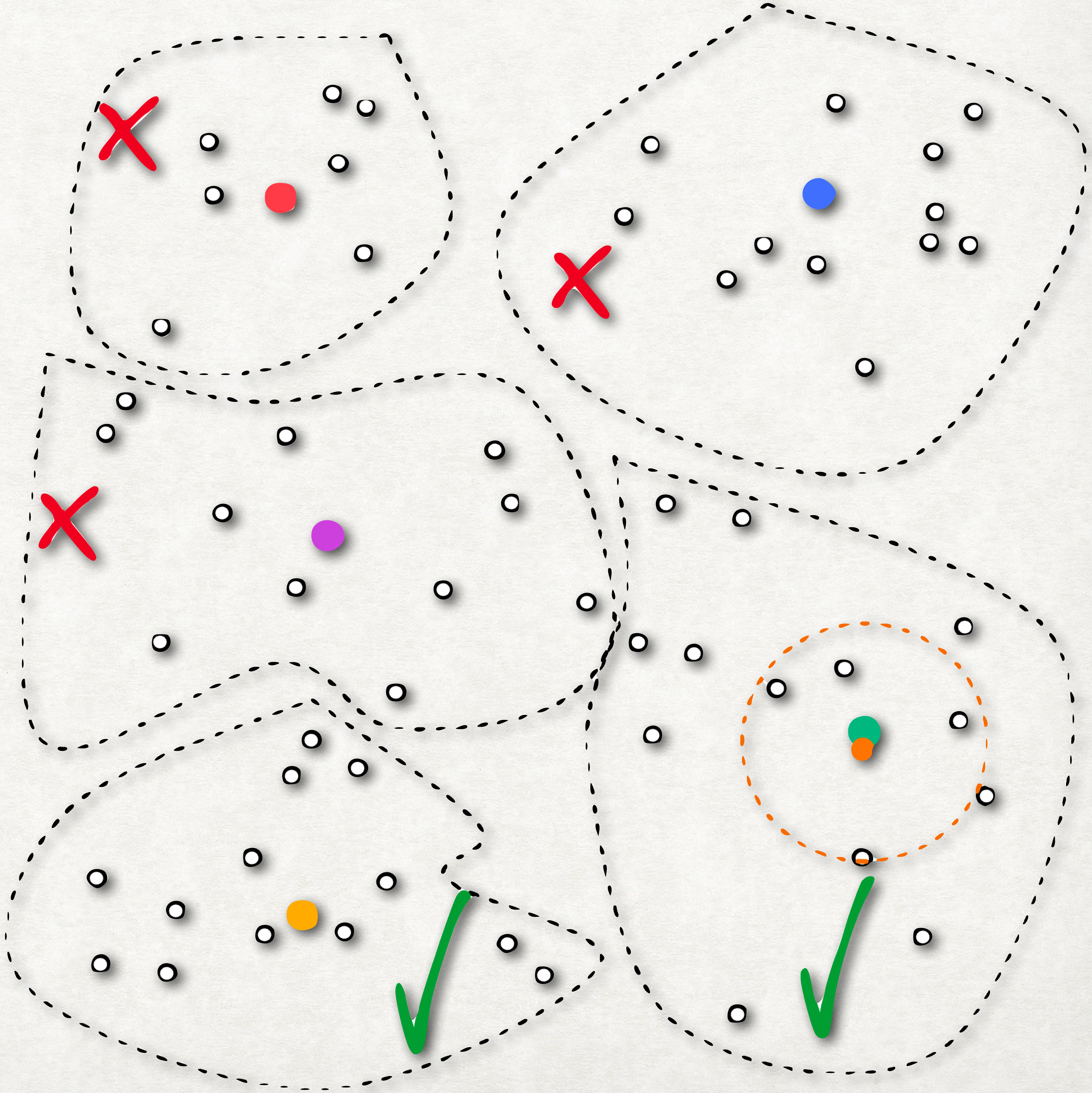
$$\min_{\mu_1,\mu_2,\ldots,\mu_k} \sum_x \min_i \|x-\mu_i\|_2$$

Every point induces a polytope in the presence of other points



Retrieval

- During search, we rank clusters by distance of query (q) from representatives (μ), then perform retrieval on the top clusters.



Inner Product

Subproblems

Compression

Indexing

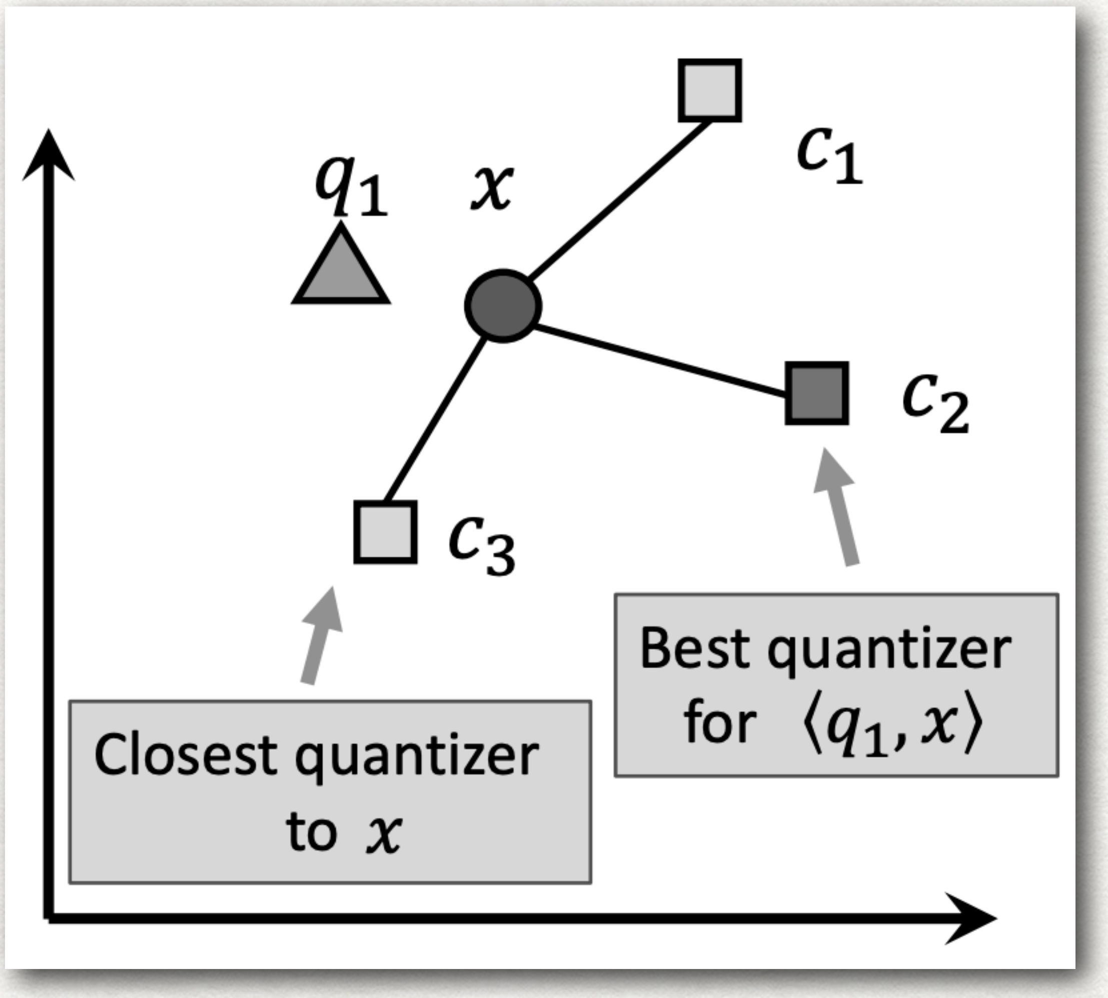
retrieval

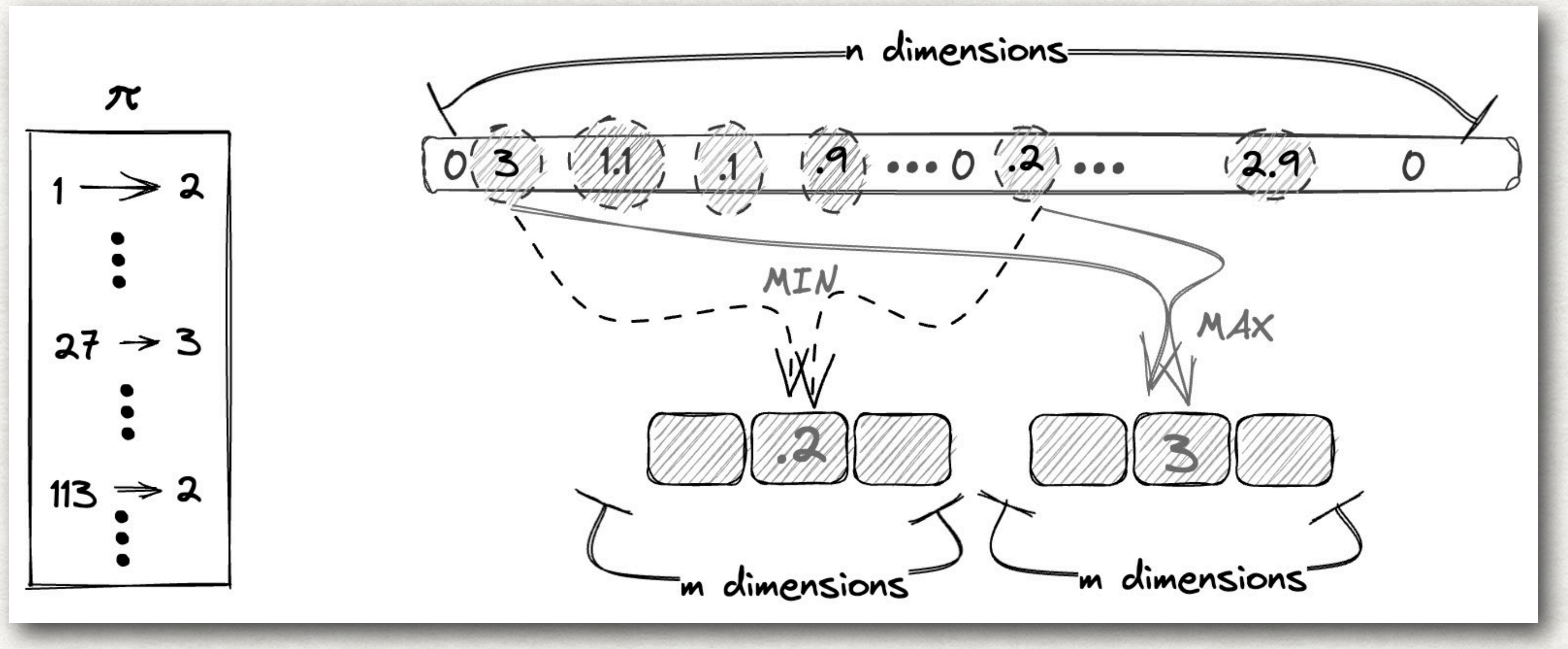
Vector Compression

- ~~**RQ**: Find a transformation $f : \mathbb{R}^N \rightarrow \mathbb{R}^n$ that preserves Euclidean distance between vectors.~~

- **RQ**: Find a transformation $f : \mathbb{R}^N \rightarrow \mathbb{R}^n$ that preserves the *order* induced by inner product of vectors:

$$f(q)^T f(x) > f(q)^T f(y) \implies q^T x > q^T y \quad w.h.p$$





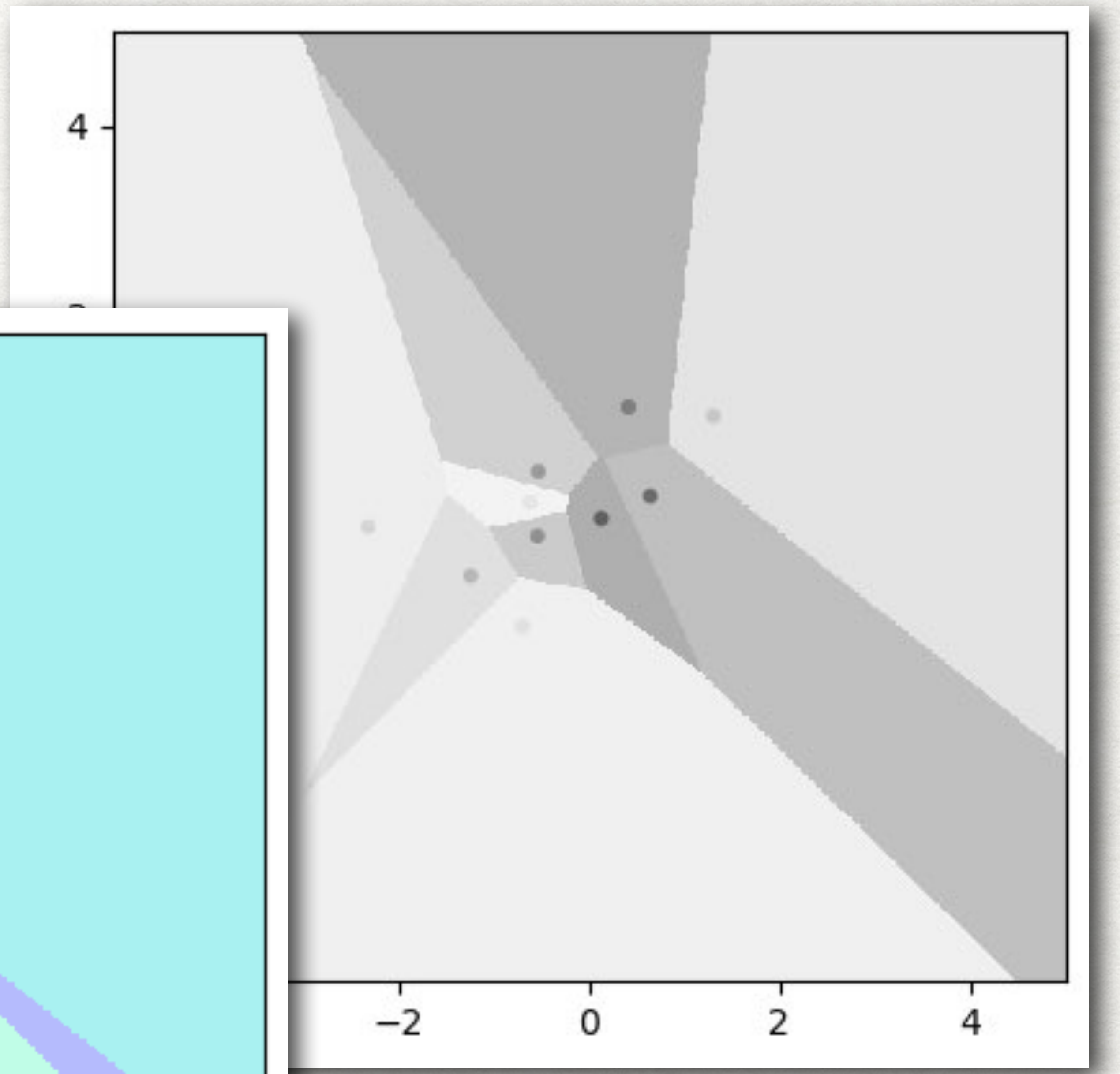
Nonlinear Sketches for Inner Product

Guo et al. "Accelerating Large-Scale Inference with Anisotropic Vector Quantization." ICML. 2020.

Bruch et al. "*An Approximate Algorithm for Maximum Inner Product Search over Streaming Sparse Vectors.*" ACM TOIS. 2023.

Polytopes and Cones

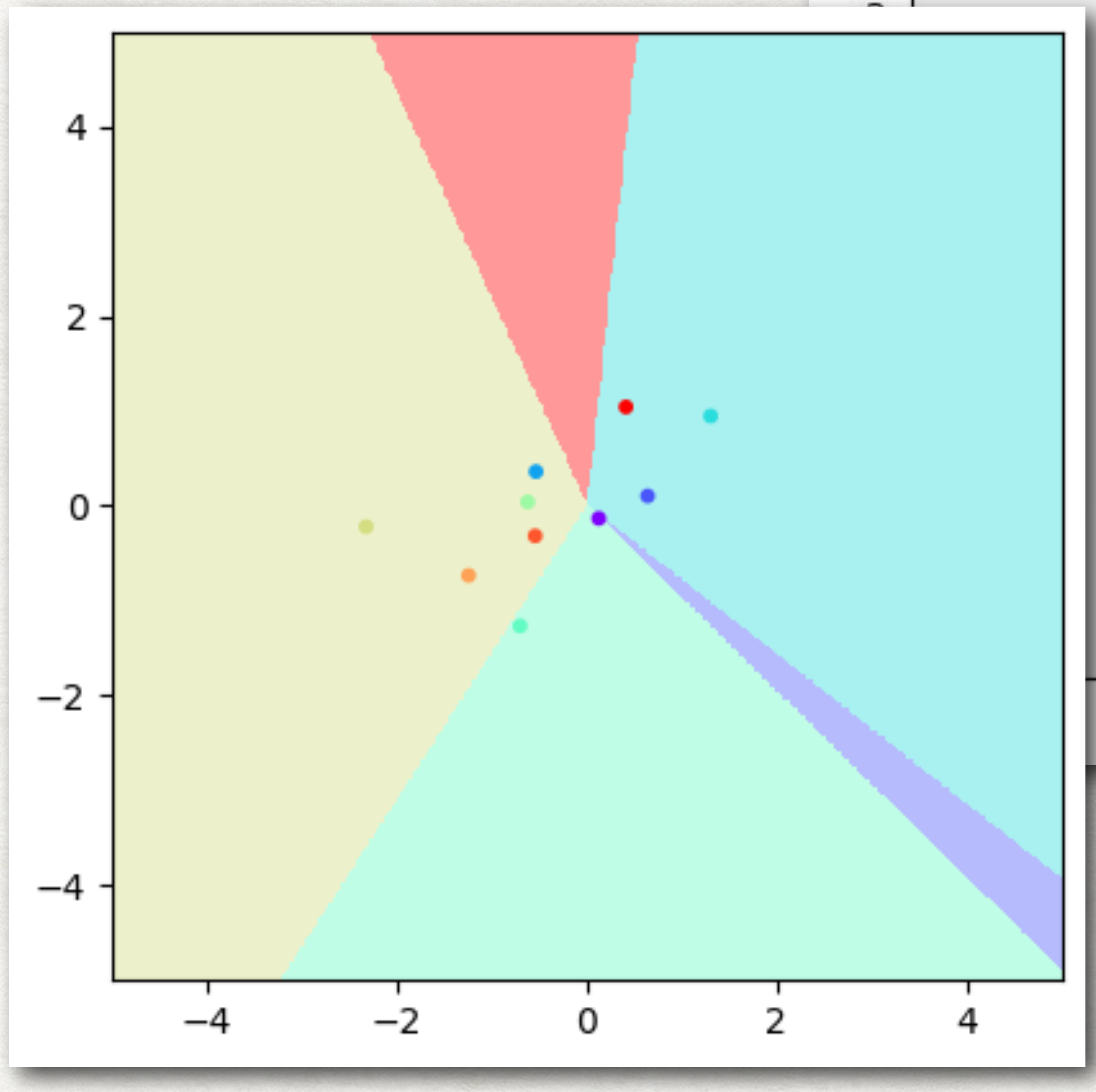
Every point induces a polytope in the presence of other points



- ~~**RQ**: Find partitions that approximate Voronoi cells.~~

- **RQ**: Find partitions that cover inner product cones.

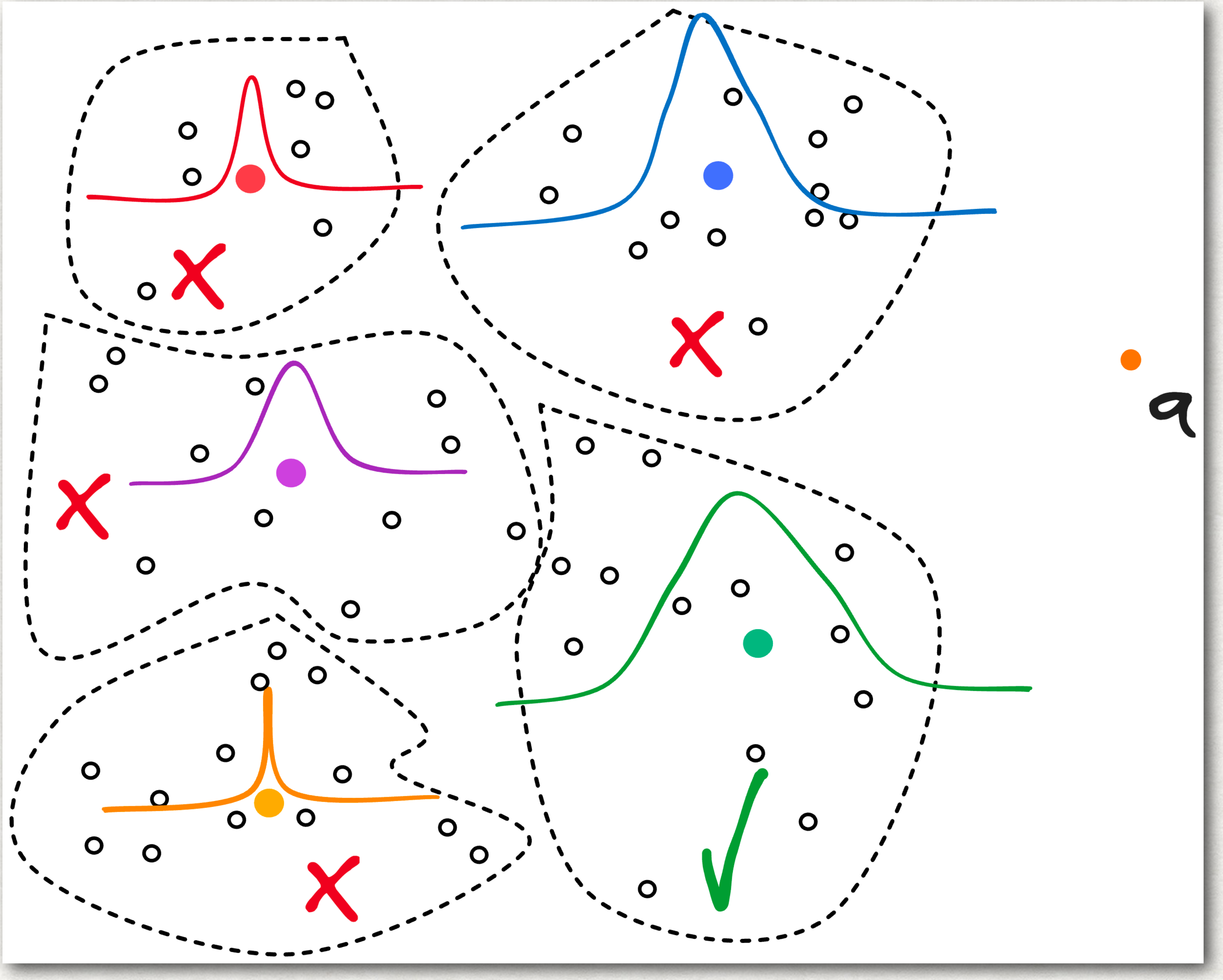
$\forall y \quad s.t. \quad x^* = argmax\, x^T y$ we have that
$\mu(x^*) = argmax\, \mu^T y \quad w.h.p$



Every point induces a convex cone (set theoretic) in the presence of other points

Retrieval

- ~~During search, we rank clusters by the inner product of query (q) with representatives (μ), then perform retrieval on the top clusters.~~
- **RQ**: Given q and a static partition of the space, rank partitions using the distribution of inner products within each partition.
  - $\mathbb{P}[|q^T X - q^T \mu| > \epsilon] < \delta$
  - Connection to online optimization (Contextual Bandits)
- **RQ**: Is space partitioning-based search sub-linear for MIPS?



Observation

Modern Information Retrieval has a variety of unique research questions that need a thorough investigation.

… and that is why I believe

Information Retrieval Needs More Theoreticians