Where Vector Search is Taking Us

HAYSTACK

Dmitry Kan

Berlin, September’22

About me

- PhD in NLP, Co-founder of Muves.io
- Senior Product Manager, Search at TomTom
- Contributor to and user of Quepid – query rating tool
- 16+ years of experience in developing search engines for start-ups and multinational technology giants
- Host of the Vector Podcast https://www.youtube.com/c/VectorPodcast
- Blogging about vector search on Medium: https://dmitry-kan.medium.com/

Outline

- Report on 1st season of Vector Podcast
- Vector Search algorithms
- Vector Search pyramid
- Use cases
- Where things are going

Launched on October 13, 2021

12 interviews

2 lectures

2 webinars

VECTOR PODCAST

by Dmitry Kan

1 LIVE recording
with 4 guests

To be continued…

Topics covered

- **Vector DBs**: Weaviate, Qdrant, Milvus, Pinecone, Vespa, Apache Solr
- **Neural search**: Jina, ZIR.AI, Haystack
- **Algorithms:** HNSW, doc2query, bi-encoders
- **Embedding layers:** Mighty
- **Sparse search & ML**

VECTOR PODCAST
Malte Pietsch
CTO at Deepset
Finding passion in NLP and bridging the gap between academia and industry with Haystack
Dmitry Kan
Host
1:26:10

Malte Pietsch - CTO, Deepset - Passion in NLP and bridgi...

Bob van Luijt (CEO, SeMI) on the Weaviate vector search...

Yury Malkov - Staff Engineer, Twitter - Author of the most...

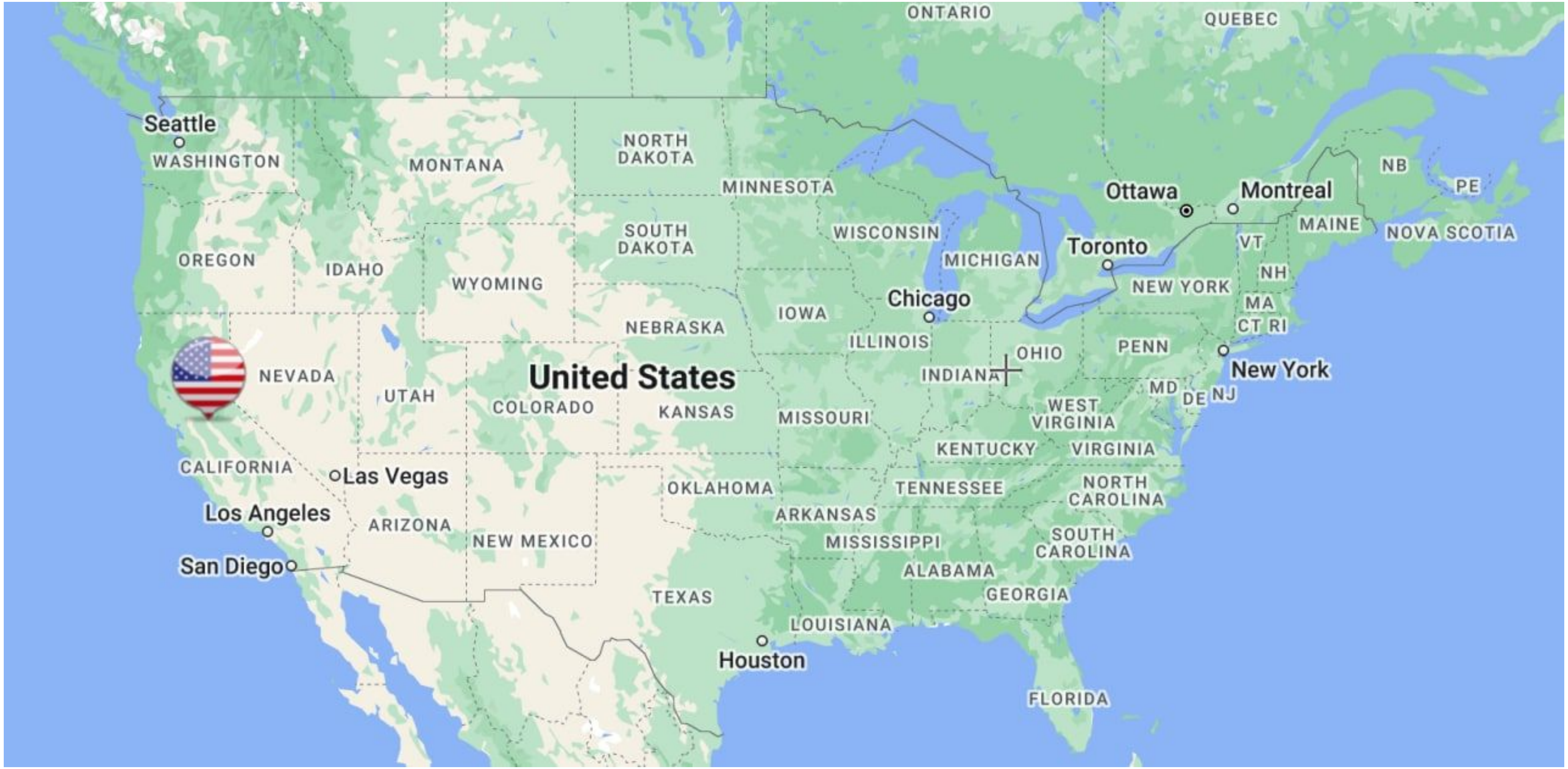
VECTOR PODCAST
Yusuf Sarıgöz
AI Research Engineer at Qdrant
Getting to know your data with metric learning
Dmitry Kan
Host
1:09:59

Yusuf Sarıgöz - AI Research

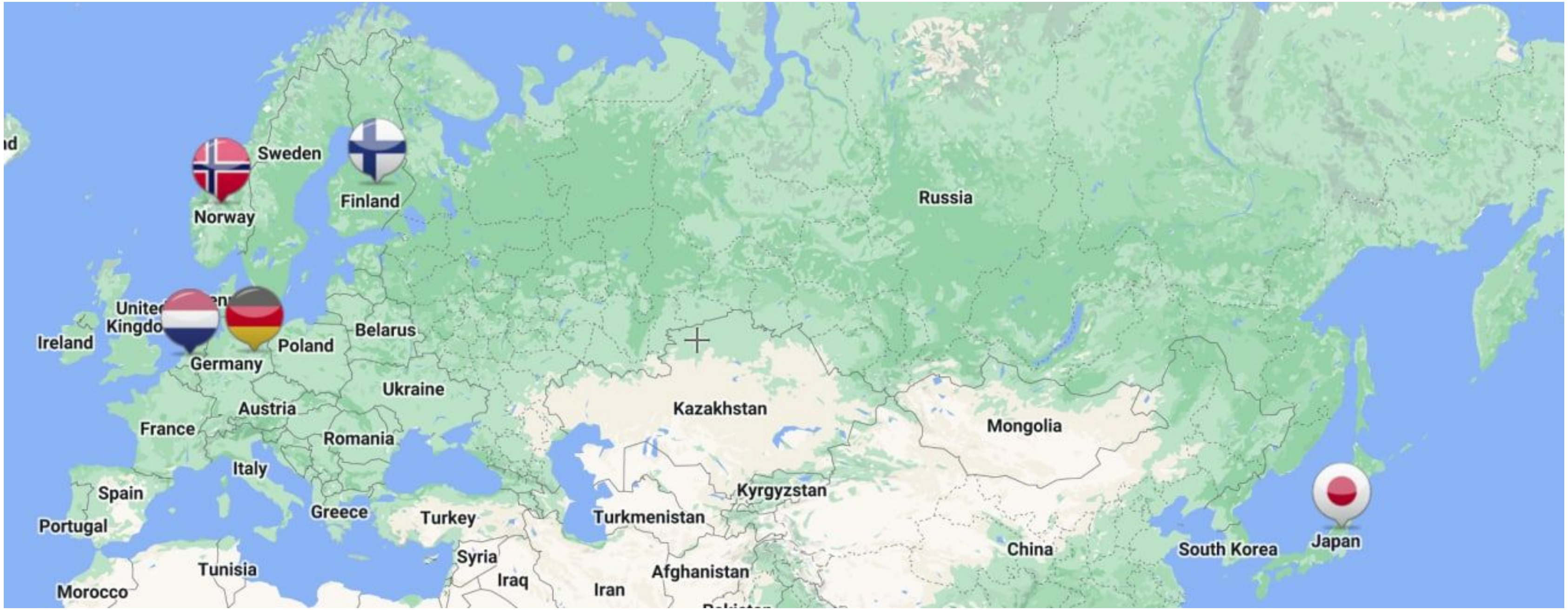
Max Irwin - Founder, MAX.IO - On economics of scale in...

Live from Berlin Buzzwords 2022 - with developers from...

Vector Search Geography: USA



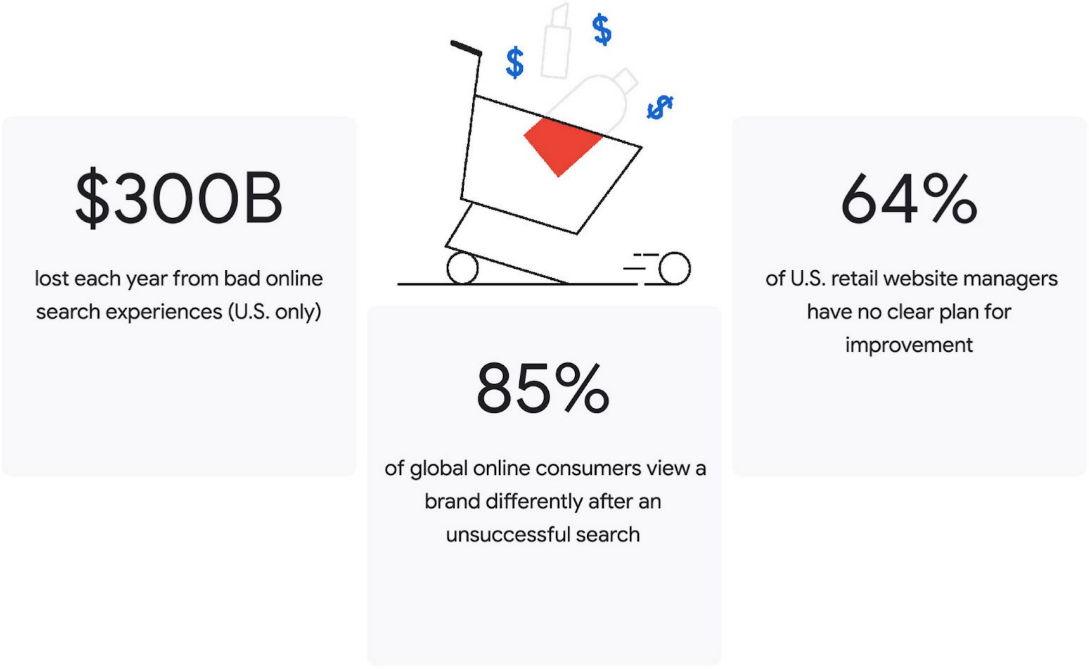
Vector Search Geography: Europe / Asia



What makers say

- Malte Pietsch (CTO Deepset, Haystack): **Metric blindness**
- Jo Bergum (Vespa, Yahoo): Keep your ears to the ground and **don’t sell hype**
- Max Irwin (Mighty): **Why** Vector Search / AI has to be locked **only** to **Python**?
- Doug Turnbull (Shopify): Stop talking about yourself as a Vector Database and switch to **Relevance oriented applications**

Search abandonment costs U.S. retailers $300 billion annually: McKinsey & Co and Google, 2021



$300B

lost each year from bad online search experiences (U.S. only)

85%

of global online consumers view a brand differently after an unsuccessful search

64%

of U.S. retail website managers have no clear plan for improvement

Source: https://cloud.google.com/blog/topics/retail/search-abandonment-impacts-retail-sales-brand-loyalty

let me tell you something about relevant
the root of the word relevant is rel

Keyword search

- *Examples: Elasticsearch, OpenSearch, Solr*
- *Relies on matching of search terms to text in an inverted index*
- *Makes it difficult to find items with similar meaning but containing different keywords*
- *Not directly suitable for multimodal or multilingual search*

EXAMPLE

Query: *A bear **eating a fish** by a river*

Result: *heron **eating a fish***

Vector search

- *Utilises neural networks models to represent objects (like text and images) and queries as high-dimensional vectors*
- *Ranking based on vector similarity*
- *Allows finding items with similar meaning or of different modality*

EXAMPLE

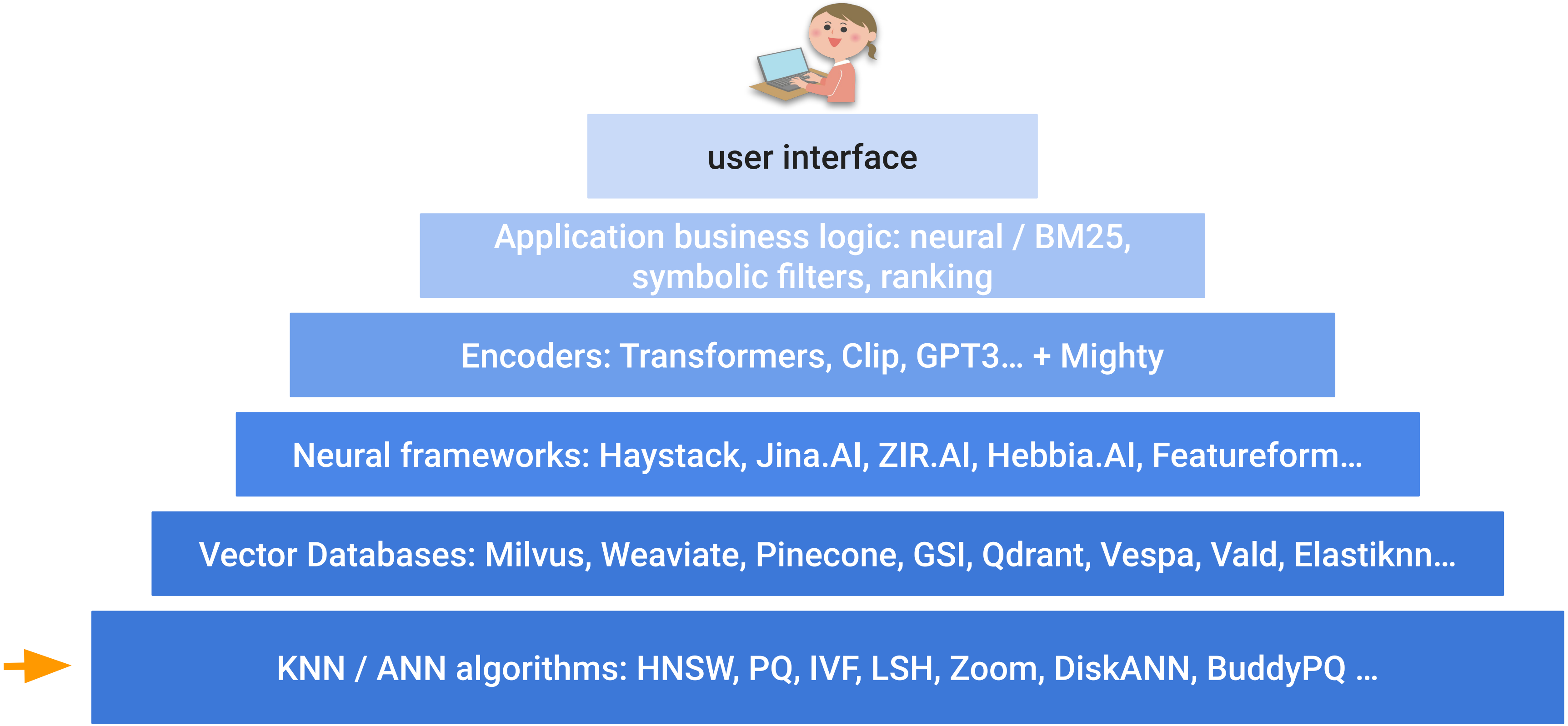
Query: *A bear eating a fish by a river*

Query vector: *[0.072893, -0.277076, 0.201384, …]*

Result vector: *[0.004142, -0.022811, 0.019714 …]*

Result:

Vector Search Pyramid



user interface

Application business logic: neural / BM25, symbolic filters, ranking

Encoders: Transformers, Clip, GPT3… + Mighty

Neural frameworks: Haystack, Jina.AI, ZIR.AI, Hebbia.AI, Featureform…

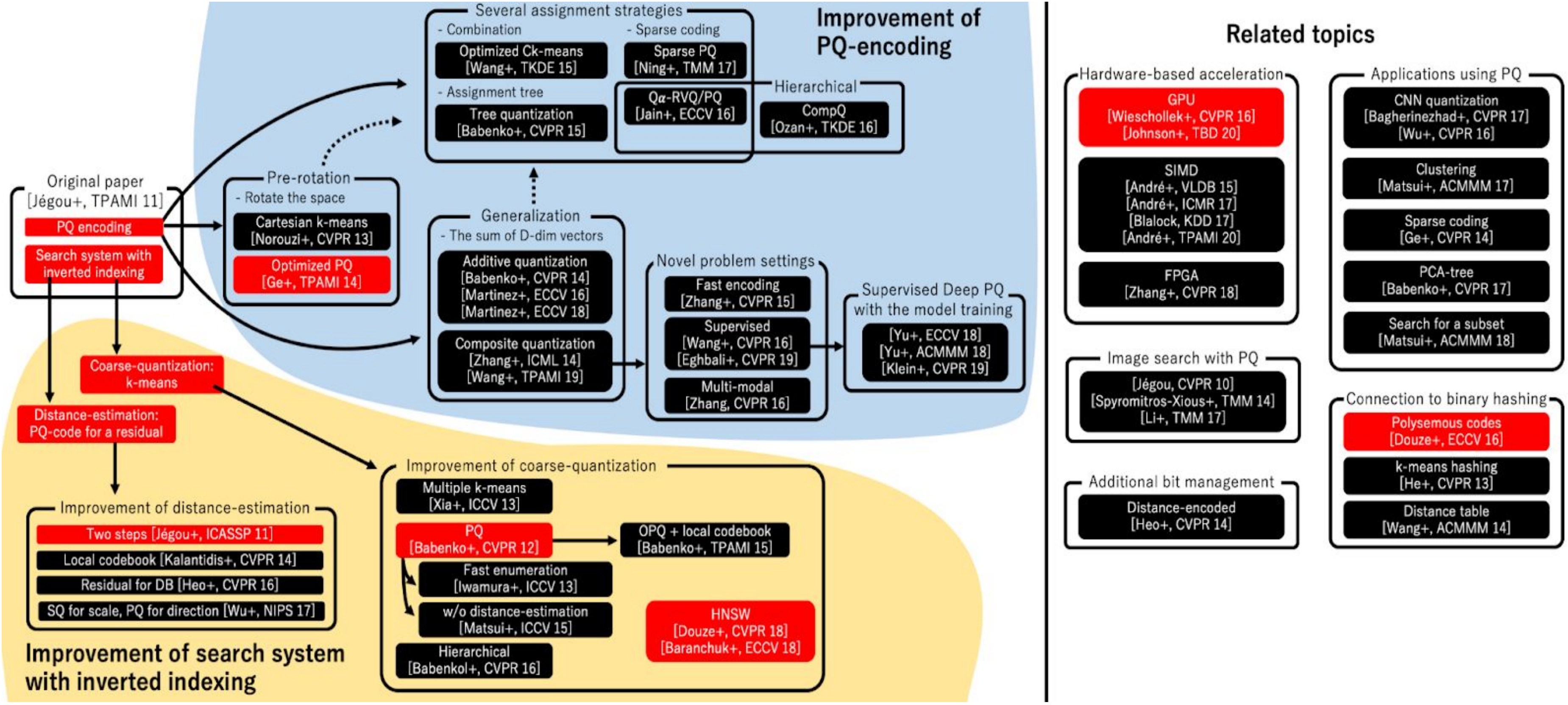
Vector Databases: Milvus, Weaviate, Pinecone, GSI, Qdrant, Vespa, Vald, Elastiknn…

KNN / ANN algorithms: HNSW, PQ, IVF, LSH, Zoom, DiskANN, BuddyPQ …

Algorithms: Big players in the game

- Spotify: ANNOY
- Microsoft (Bing team): Zoom, DiskANN, SPTAG
  - Azure Cognitive Search
- Amazon: KNN based on HNSW in OpenSearch
- Google: ScaNN
- Yahoo! Japan: NGT
- Facebook: FAISS, PQ (CPU & GPU)
- Baidu: IPDG (Baidu Cloud)
- Yandex
- NVIDIA
- Intel

ANN algorithms



Original paper [Jégou+, TPAMI 11]

PQ encoding

Search system with inverted indexing

Improvement of PQ-encoding

Several assignment strategies

- Combination

Optimized Ck-means [Wang+, TKDE 15]

- Sparse coding

Sparse PQ [Ning+, TMM 17]

- Assignment tree

Tree quantization [Babenko+, CVPR 15]

Qα-RVQ/PQ [Jain+, ECCV 16]

Hierarchical

CompQ [Ozan+, TKDE 16]

Pre-rotation

- Rotate the space

Cartesian k-means [Norouzi+, CVPR 13]

Optimized PQ [Ge+, TPAMI 14]

Generalization

- The sum of D-dim vectors

Additive quantization [Babenko+, CVPR 14] [Martinez+, ECCV 16] [Martinez+, ECCV 18]

Composite quantization [Zhang+, ICML 14] [Wang+, TPAMI 19]

Novel problem settings

Fast encoding [Zhang+, CVPR 15]

Supervised [Wang+, CVPR 16] [Eghbali+, CVPR 19]

Multi-modal [Zhang, CVPR 16]

Supervised Deep PQ with the model training

[Yu+, ECCV 18] [Yu+, ACMMM 18] [Klein+, CVPR 19]

Coarse-quantization: k-means

Distance-estimation: PQ-code for a residual

Improvement of distance-estimation

Two steps [Jégou+, ICASSP 11]

Local codebook [Kalantidis+, CVPR 14]

Residual for DB [Heo+, CVPR 16]

SQ for scale, PQ for direction [Wu+, NIPS 17]

Improvement of search system with inverted indexing

Improvement of coarse-quantization

Multiple k-means [Xia+, ICCV 13]

PQ [Babenko+, CVPR 12]

OPQ + local codebook [Babenko+, TPAMI 15]

Fast enumeration [Iwamura+, ICCV 13]

w/o distance-estimation [Matsui+, ICCV 15]

Hierarchical [Babenkol+, CVPR 16]

HNSW [Douze+, CVPR 18] [Baranchuk+, ECCV 18]

Related topics

Hardware-based acceleration

GPU [Wieschollek+, CVPR 16] [Johnson+, TBD 20]

SIMD [André+, VLDB 15] [André+, ICMR 17] [Blalock, KDD 17] [André+, TPAMI 20]

FPGA [Zhang+, CVPR 18]

Image search with PQ

[Jégou, CVPR 10] [Spyromitros-Xious+, TMM 14] [Li+, TMM 17]

Additional bit management

Distance-encoded [Heo+, CVPR 14]

Applications using PQ

CNN quantization [Bagherinezhad+, CVPR 17] [Wu+, CVPR 16]

Clustering [Matsui+, ACMMM 17]

Sparse coding [Ge+, CVPR 14]

PCA-tree [Babenko+, CVPR 17]

Search for a subset [Matsui+, ACMMM 18]

Connection to binary hashing

Polysemous codes [Douze+, ECCV 16]

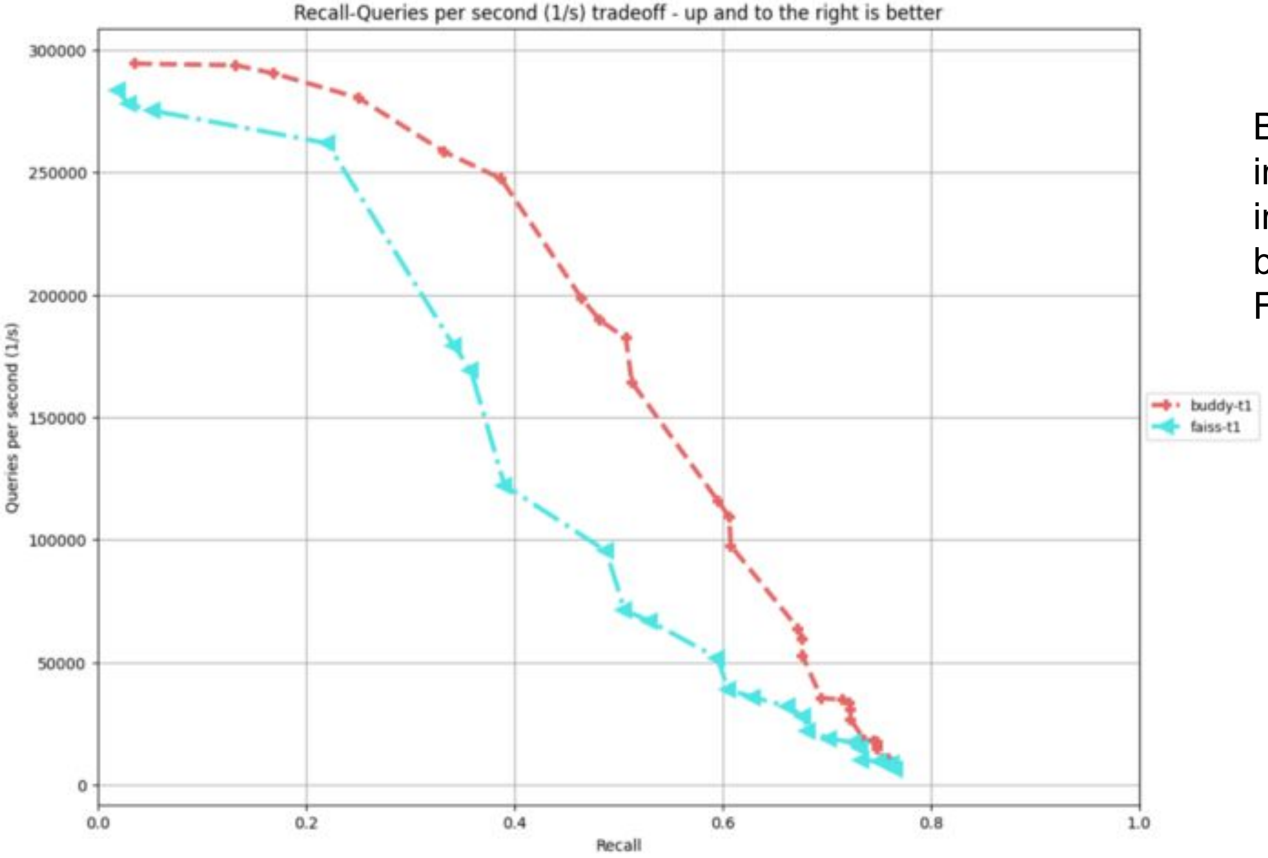
k-means hashing [He+, CVPR 13]

Distance table [Wang+, ACMMM 14]

BigANN Competition @ NeurIPS’21

- Max Irwin
- Alex Semenov
- Aarne Talman
- Leo Joffe
- Alex Klibisz
- Dmitry Kan

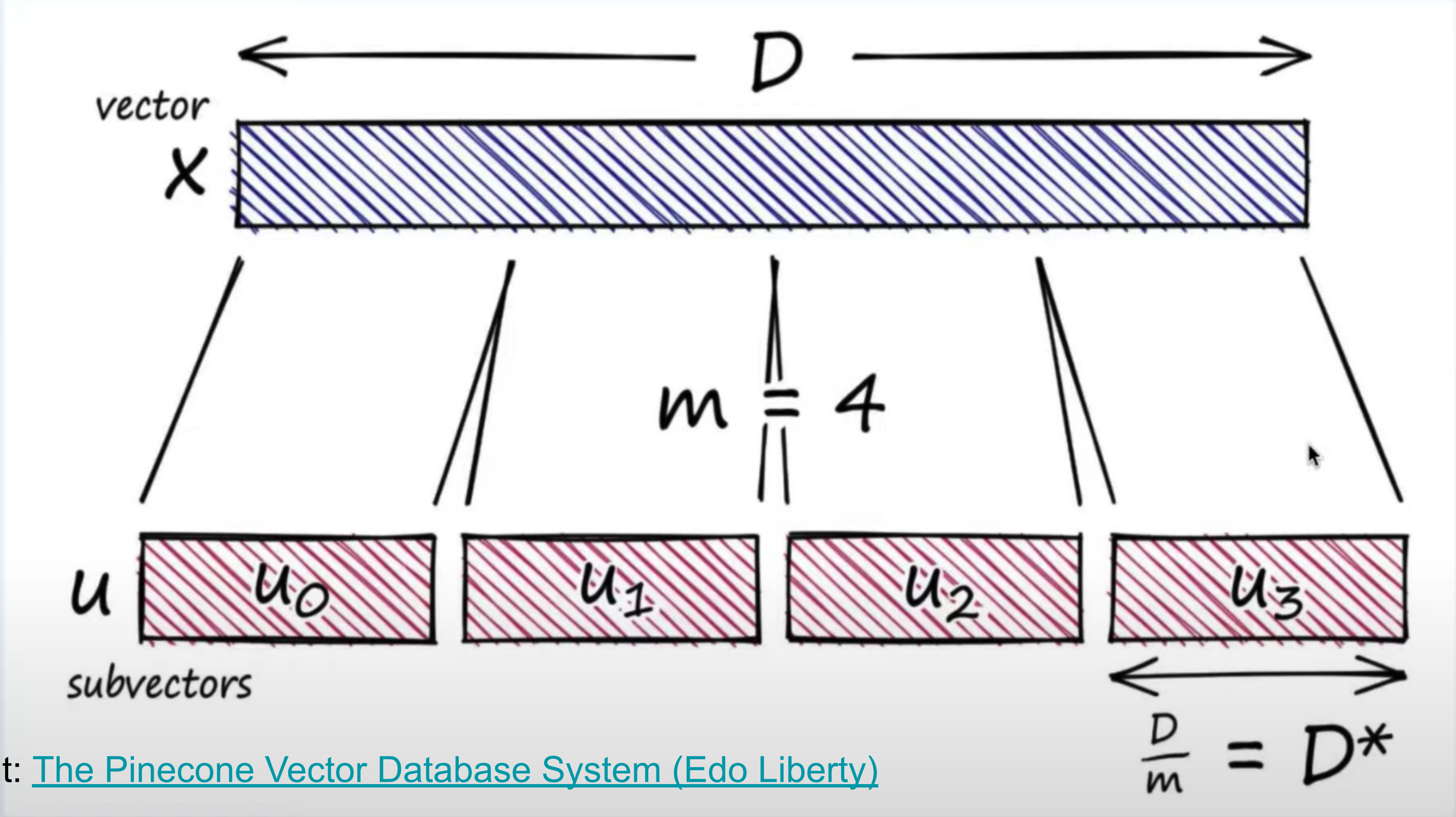
Billion-Scale ANN Algorithm Challenge



BuddyPQ increases **12%** in Recall over baseline FAISS

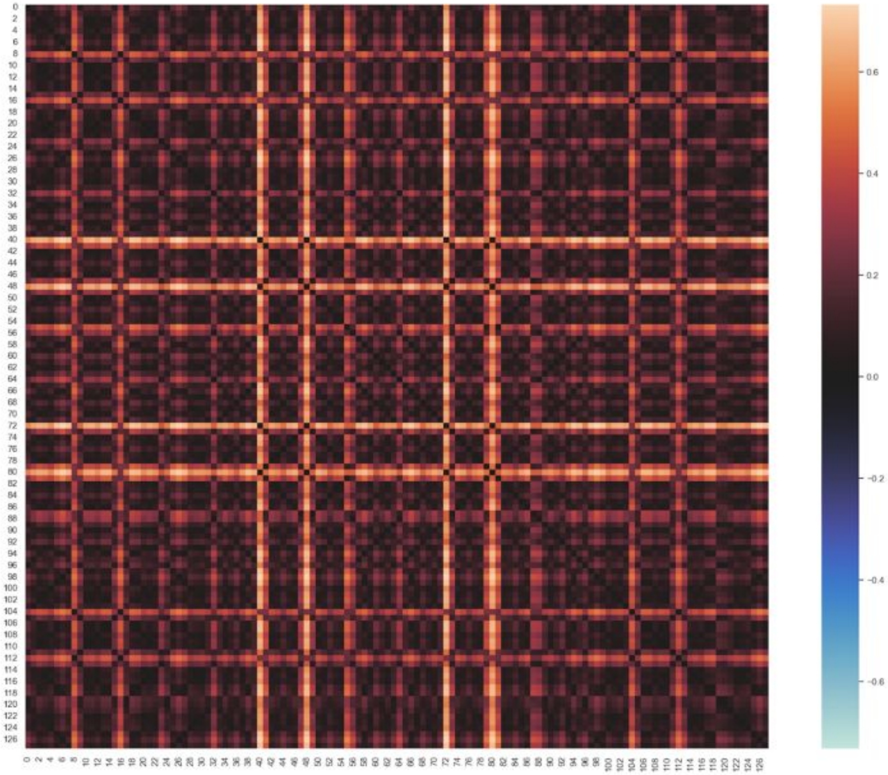
https://bit.ly/3ApqYYQ

PQ (Product Quantization)

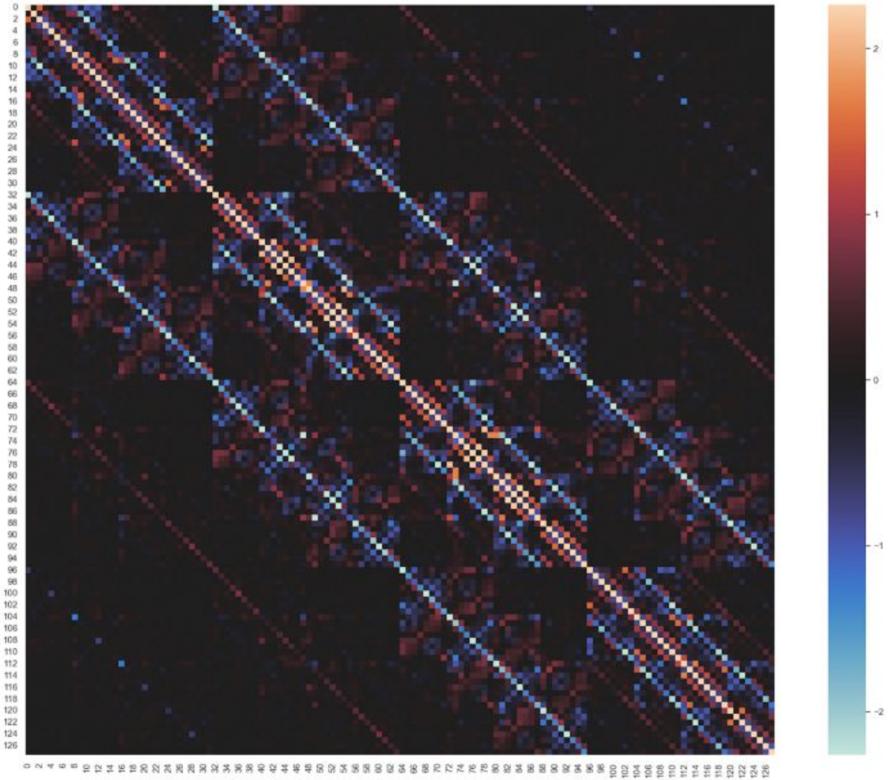


Credit: The Pinecone Vector Database System (Edo Liberty)

BuddyPQ: improving 12% recall over FAISS



BIGANN dataset Kolmogorov-Smirnov dimension test matrix for the first 100000 points. A higher number indicates a less similar distribution



Variance Inflation Factor (Multicollinearity) (~2.25)

Product Quantization vs Stemming

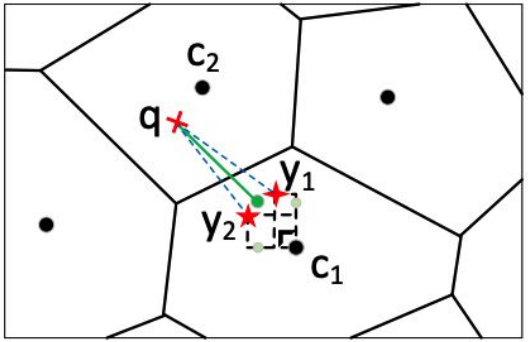
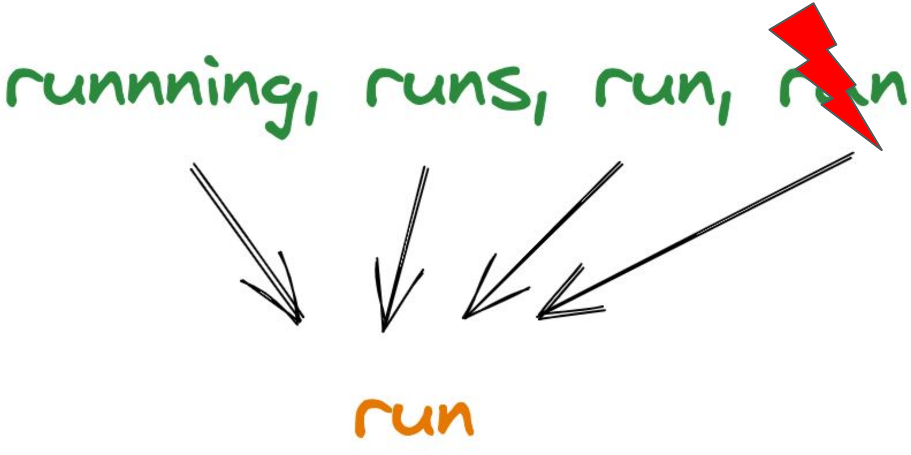


Figure 2: 2D IVFPQ example of the lossy procedure of quantization-based approaches. q denotes the query, y_1 and y_2 (drawn as a red cross) denotes vectors. c_1 and c_2 represent cluster centroids.



Motivation for vector search

- Text search can be *solved** with inverted index and BM25 or TF-IDF ranking
- But what if your data is images, audio, video?
- Can you find images with textual queries?

Motivation for vector search

With inverted index you can represent documents as a sparse vector:

[1, 1, 0, 1, 0, 0, 0, 0, ..., 0]

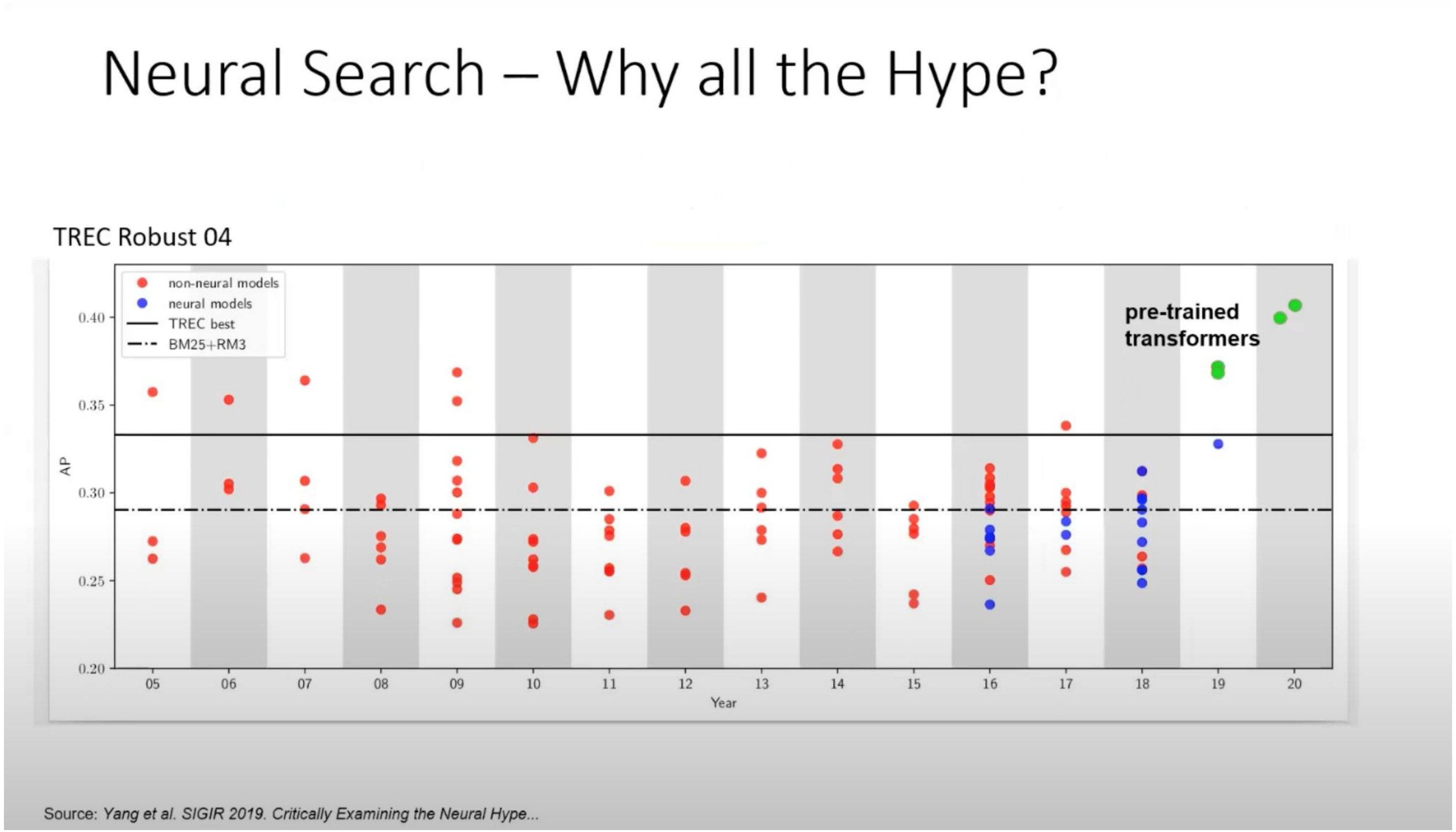
Problem 1: this bag-of-words representation does not take into account semantic context in which query words appear.

“Capital” as a main city OR as some monetary value?

Problem 2: Does not respect word order:

Visa from Finland to Canada

Visa from Canada to Finland



Neural Search – Why all the Hype?

TREC Robust 04

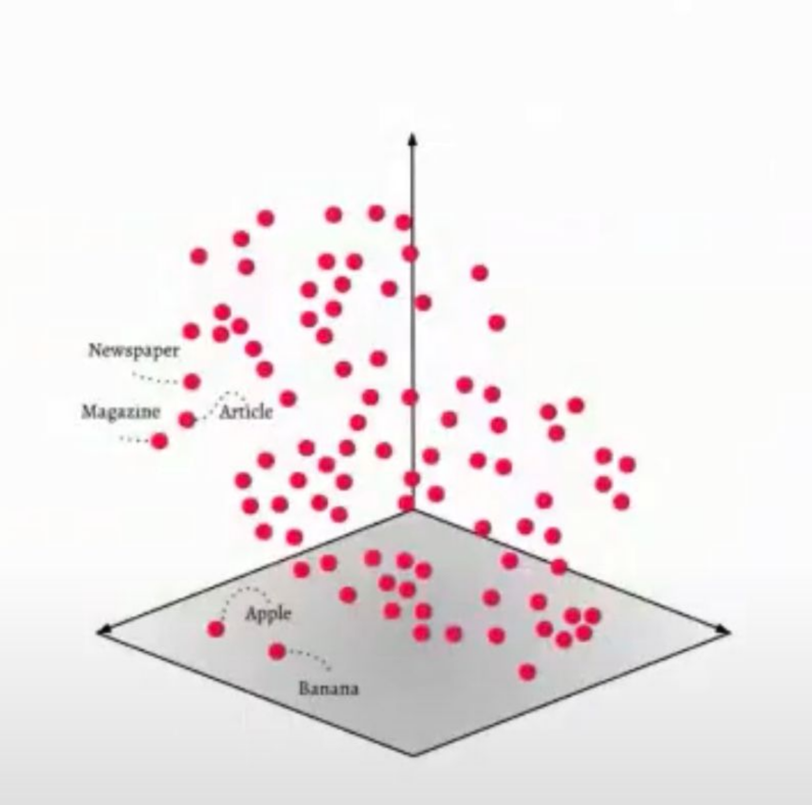
Source: *Yang et al. SIGIR 2019. Critically Examining the Neural Hype...*

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q

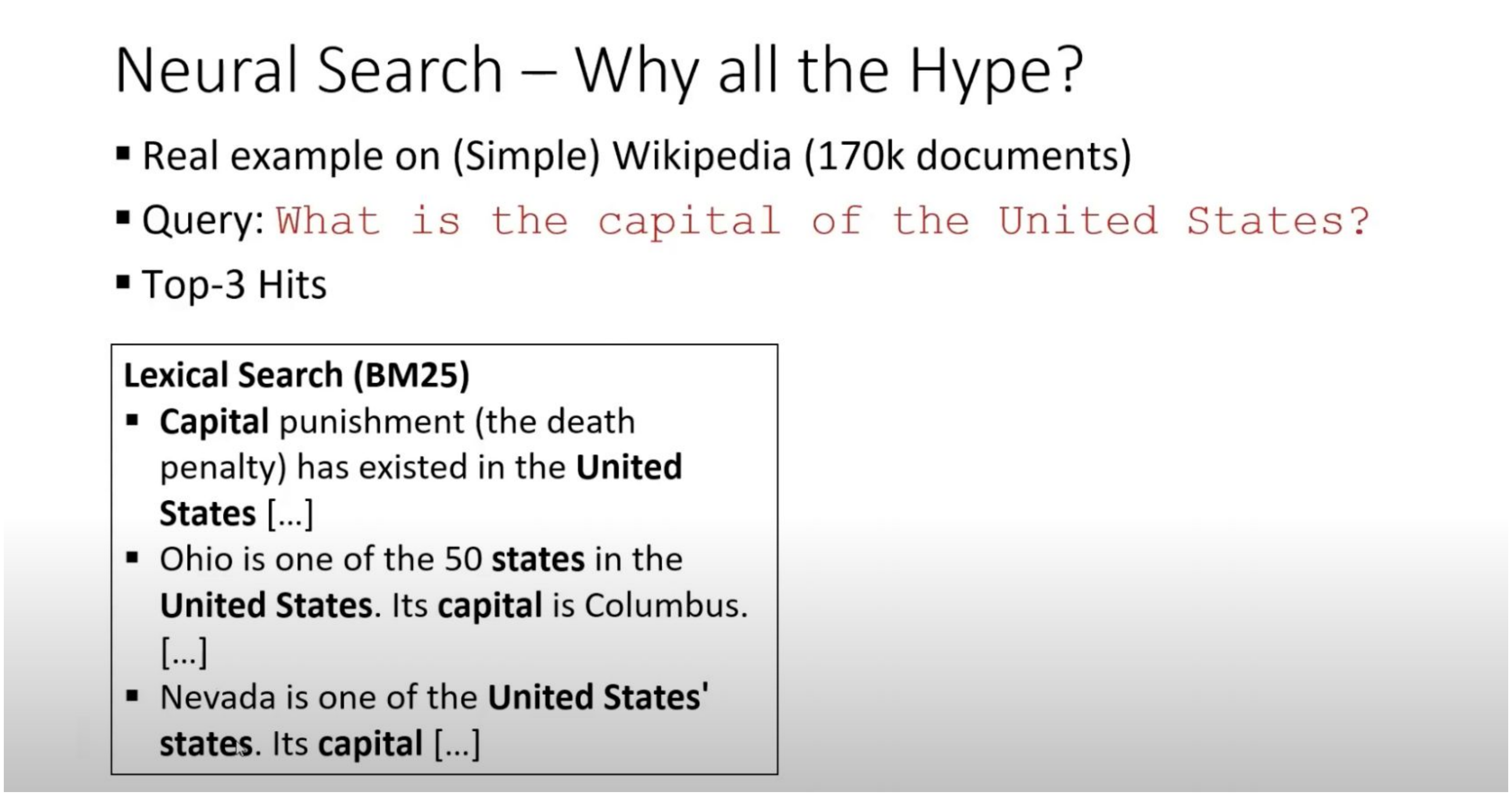
Vector search in a nutshell

Vector (=Neural) search is

- a way to **represent** and
- **search** your objects (documents, songs, images) in a geometric space
- usually of high-dimension in the form of a **dense** embedding
- a vector of numbers: [0.9, -0.1, 0.15, ...]



Credit: Weaviate V1.0 release - virtual meetup



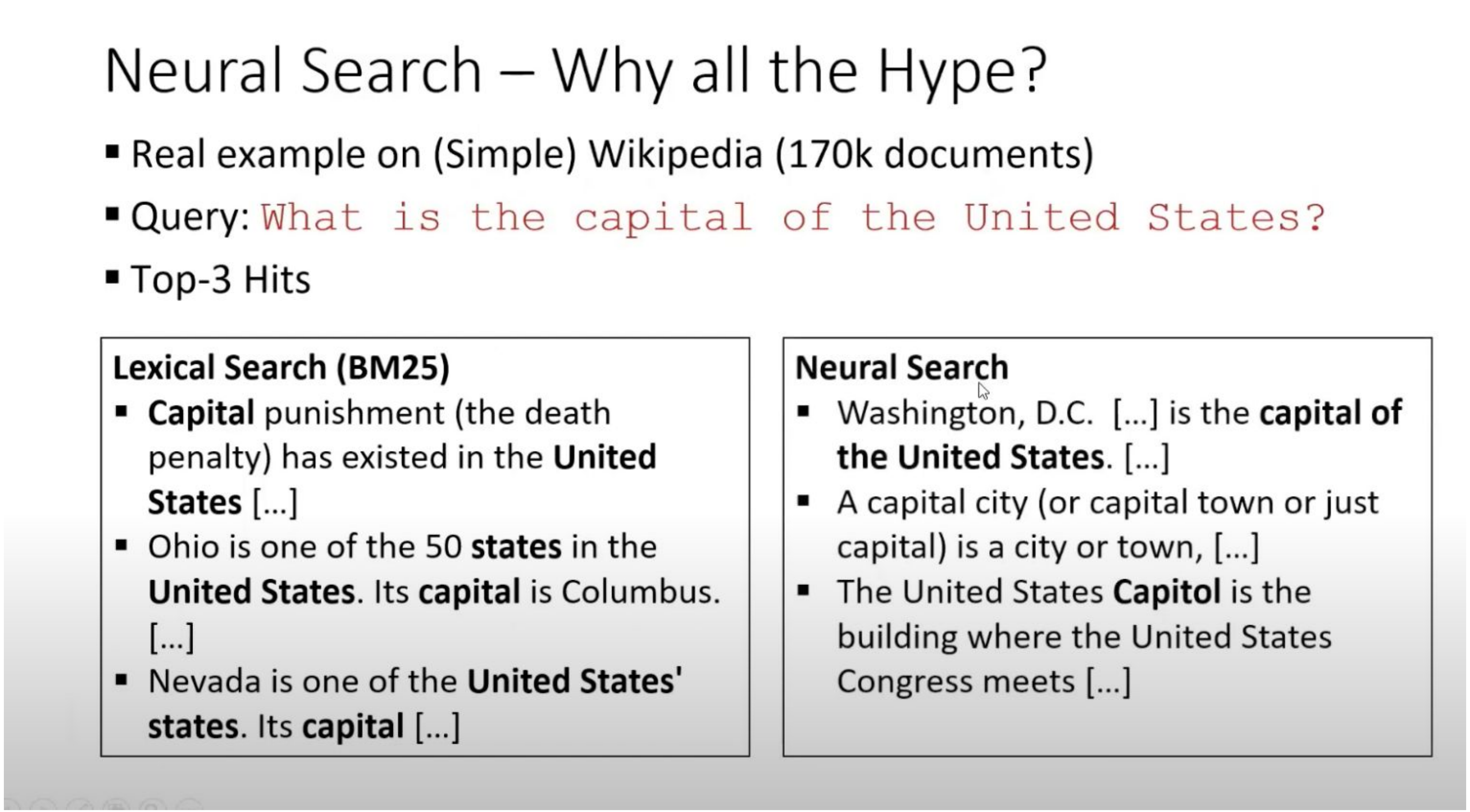
Neural Search – Why all the Hype?

- Real example on (Simple) Wikipedia (170k documents)
- Query: `What is the capital of the United States?`
- Top-3 Hits

Lexical Search (BM25)

- **Capital** punishment (the death penalty) has existed in the **United States** […]
- Ohio is one of the 50 **states** in the **United States**. Its **capital** is Columbus. […]
- Nevada is one of the **United States'** **states**. Its **capital** […]

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q



Neural Search – Why all the Hype?

- Real example on (Simple) Wikipedia (170k documents)
- Query: `What is the capital of the United States?`
- Top-3 Hits

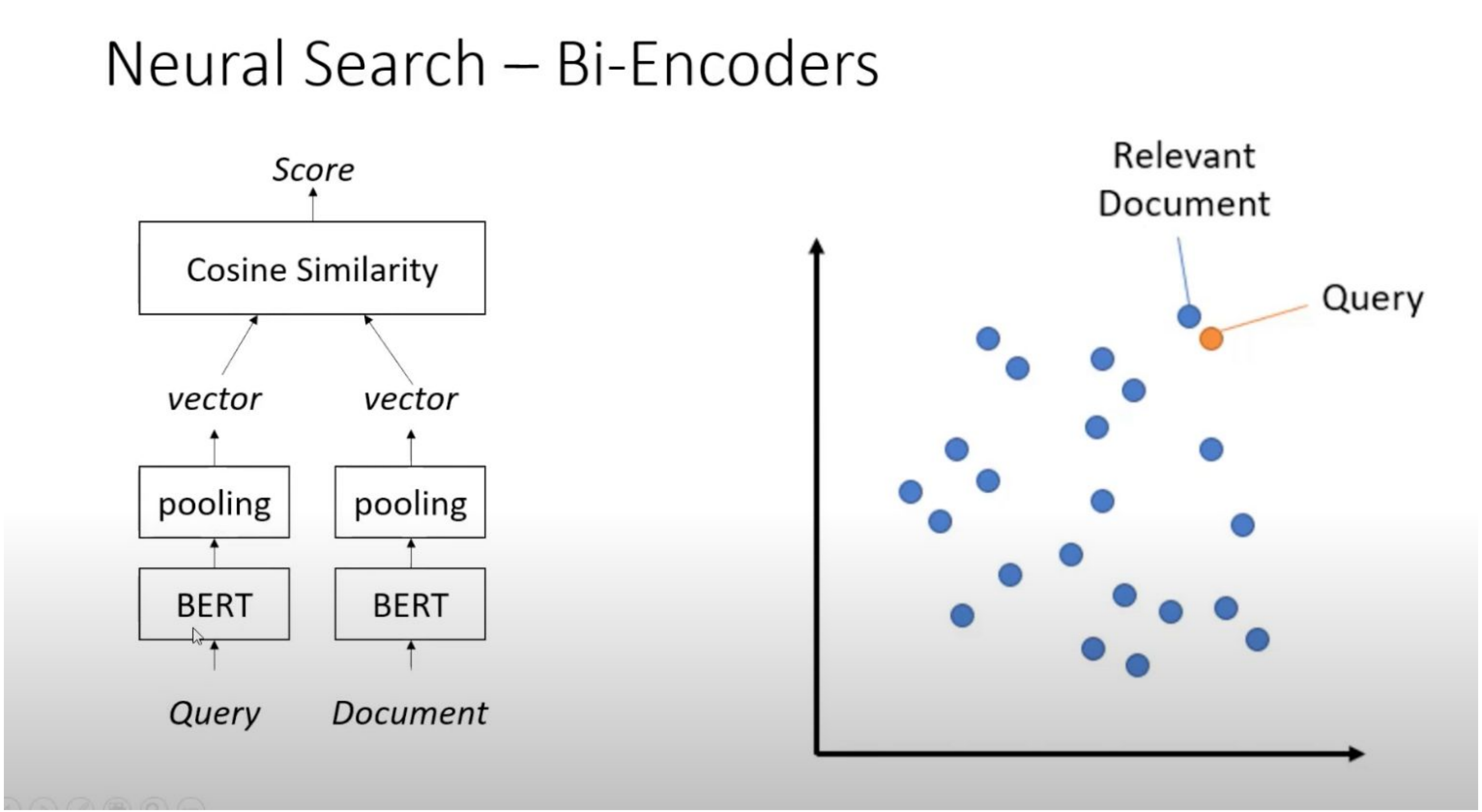
Lexical Search (BM25)

- **Capital** punishment (the death penalty) has existed in the **United States** […]
- Ohio is one of the 50 **states** in the **United States**. Its **capital** is Columbus. […]
- Nevada is one of the **United States' states**. Its **capital** […]

Neural Search

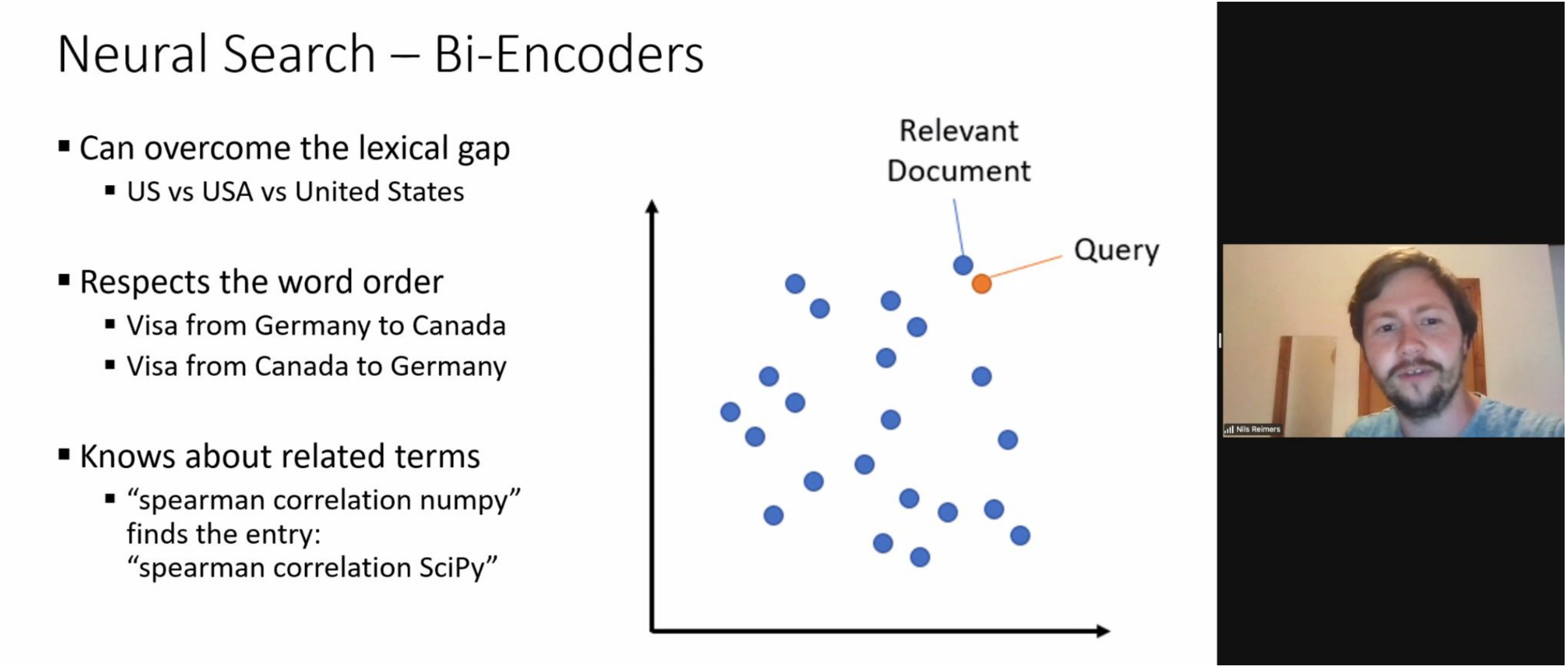
- Washington, D.C. […] is the **capital of the United States**. […]
- A capital city (or capital town or just capital) is a city or town, […]
- The United States **Capitol** is the building where the United States Congress meets […]

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q



Neural Search – Bi-Encoders

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q

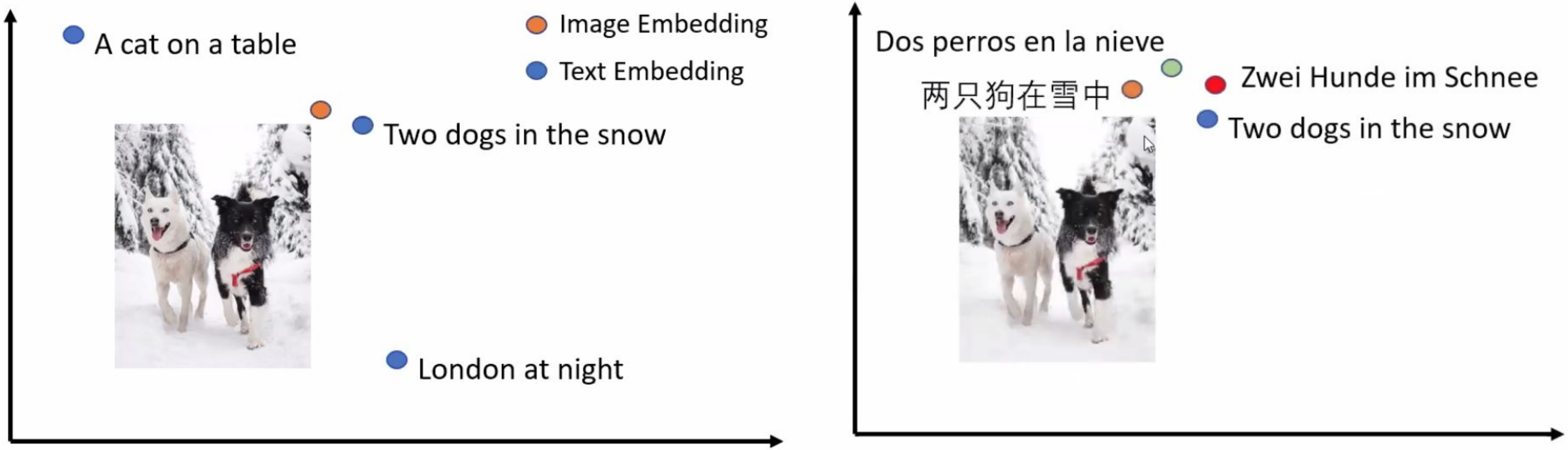


Neural Search – Bi-Encoders

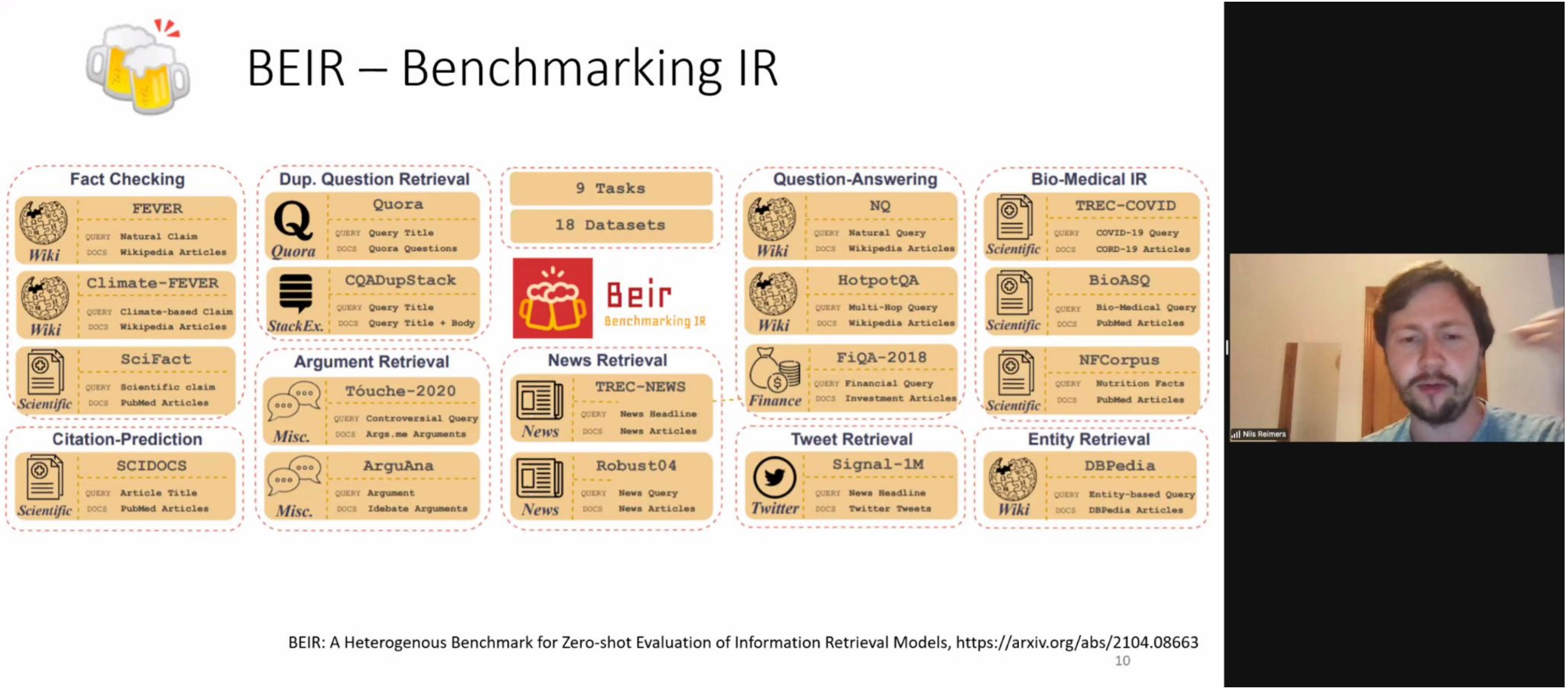
- Can overcome the lexical gap
  - US vs USA vs United States
- Respects the word order
  - Visa from Germany to Canada
  - Visa from Canada to Germany
- Knows about related terms
  - “spearman correlation numpy” finds the entry: “spearman correlation SciPy”

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q

Multi-Modal & Multi-Lingual Search



Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q



BEIR – Benchmarking IR

Fact Checking

- FEVER (Wiki) — QUERY Natural Claim; DOCS Wikipedia Articles
- Climate-FEVER (Wiki) — QUERY Climate-based Claim; DOCS Wikipedia Articles
- SciFact (Scientific) — QUERY Scientific claim; DOCS PubMed Articles

Citation-Prediction

- SCIDOCS (Scientific) — QUERY Article Title; DOCS PubMed Articles

Dup. Question Retrieval

- Quora (Quora) — QUERY Query Title; DOCS Quora Questions
- CQADupStack (StackEx.) — QUERY Query Title; DOCS Query Title + Body

Argument Retrieval

- Touché-2020 (Misc.) — QUERY Controversial Query; DOCS Args.me Arguments
- ArguAna (Misc.) — QUERY Argument; DOCS Idebate Arguments

9 Tasks

18 Datasets

Beir
Benchmarking IR

News Retrieval

- TREC-NEWS (News) — QUERY News Headline; DOCS News Articles
- Robust04 (News) — QUERY News Query; DOCS News Articles

Question-Answering

- NQ (Wiki) — QUERY Natural Query; DOCS Wikipedia Articles
- HotpotQA (Wiki) — QUERY Multi-Hop Query; DOCS Wikipedia Articles
- FiQA-2018 (Finance) — QUERY Financial Query; DOCS Investment Articles

Tweet Retrieval

- Signal-1M (Twitter) — QUERY News Headline; DOCS Twitter Tweets

Bio-Medical IR

- TREC-COVID (Scientific) — QUERY COVID-19 Query; DOCS CORD-19 Articles
- BioASQ (Scientific) — QUERY Bio-Medical Query; DOCS PubMed Articles
- NFCorpus (Scientific) — QUERY Nutrition Facts; DOCS PubMed Articles

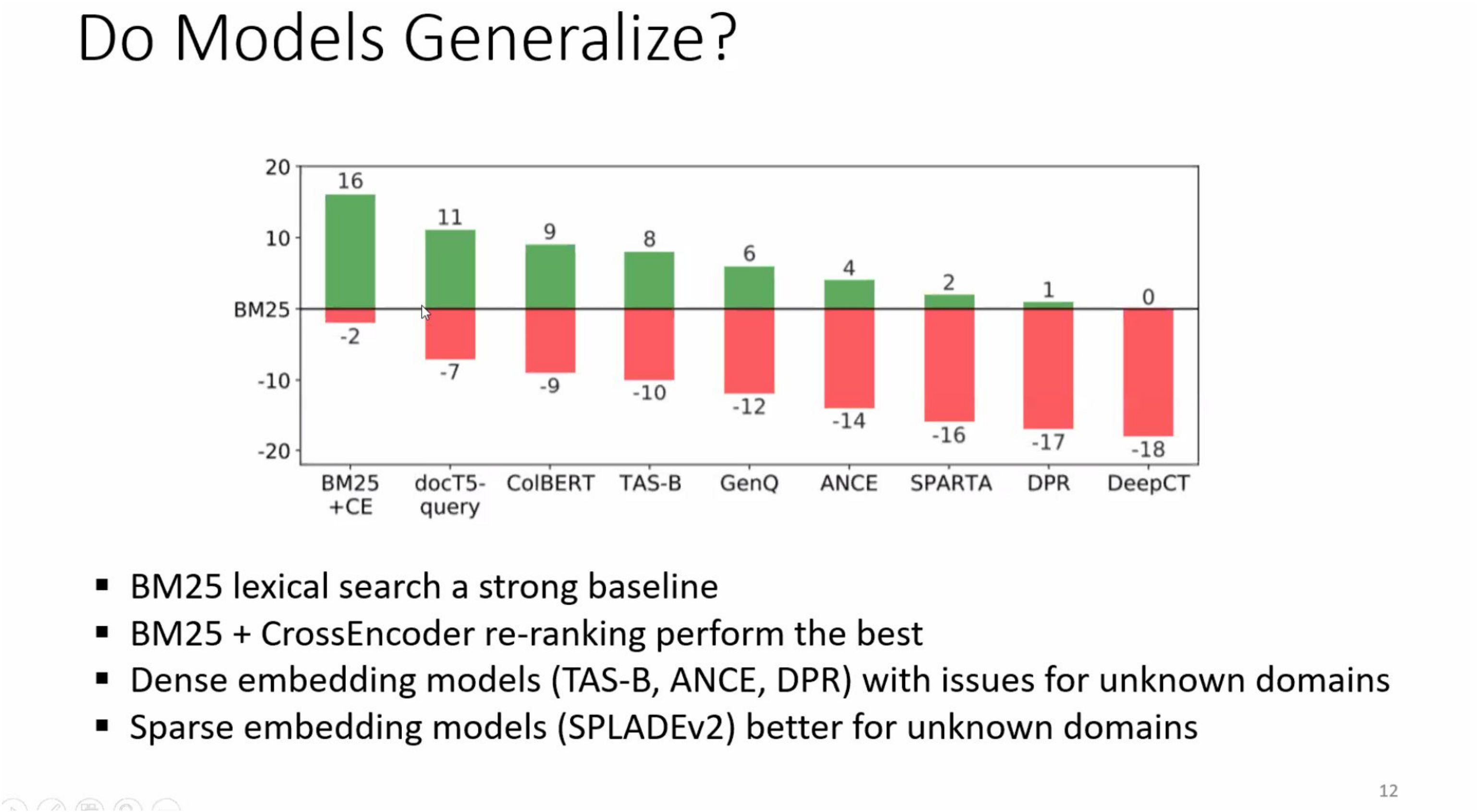
Entity Retrieval

- DBPedia (Wiki) — QUERY Entity-based Query; DOCS DBPedia Articles

BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, https://arxiv.org/abs/2104.08663

10

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q

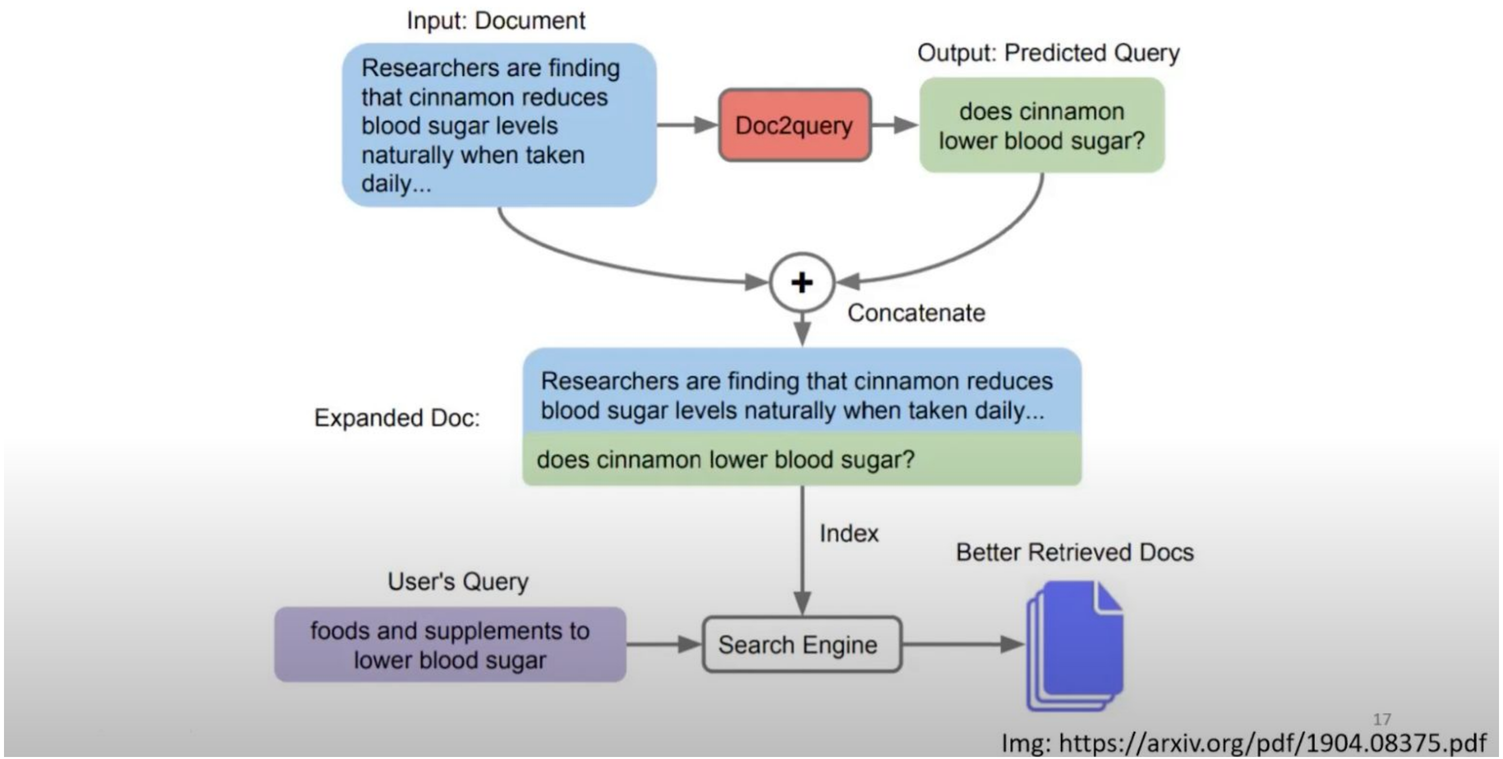


Do Models Generalize?

- BM25 lexical search a strong baseline
- BM25 + CrossEncoder re-ranking perform the best
- Dense embedding models (TAS-B, ANCE, DPR) with issues for unknown domains
- Sparse embedding models (SPLADEv2) better for unknown domains

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q

doc2Query



Input: Document

Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...

Doc2query

Output: Predicted Query

does cinnamon lower blood sugar?

Concatenate

Expanded Doc:

Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...

does cinnamon lower blood sugar?

Index

User's Query

foods and supplements to lower blood sugar

Search Engine

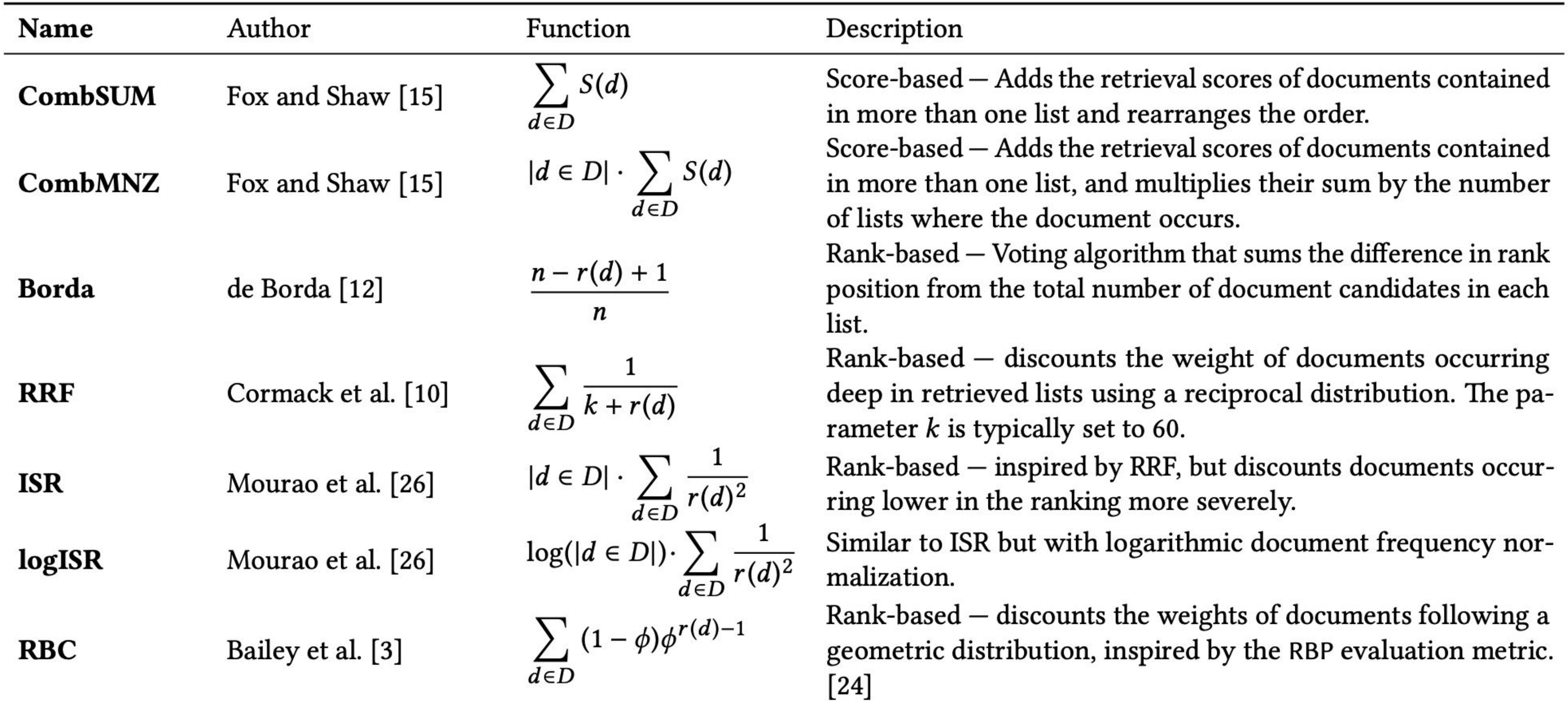
Better Retrieved Docs

17

Img: https://arxiv.org/pdf/1904.08375.pdf

Credit: https://www.youtube.com/watch?v=qzQPbIcQu9Q

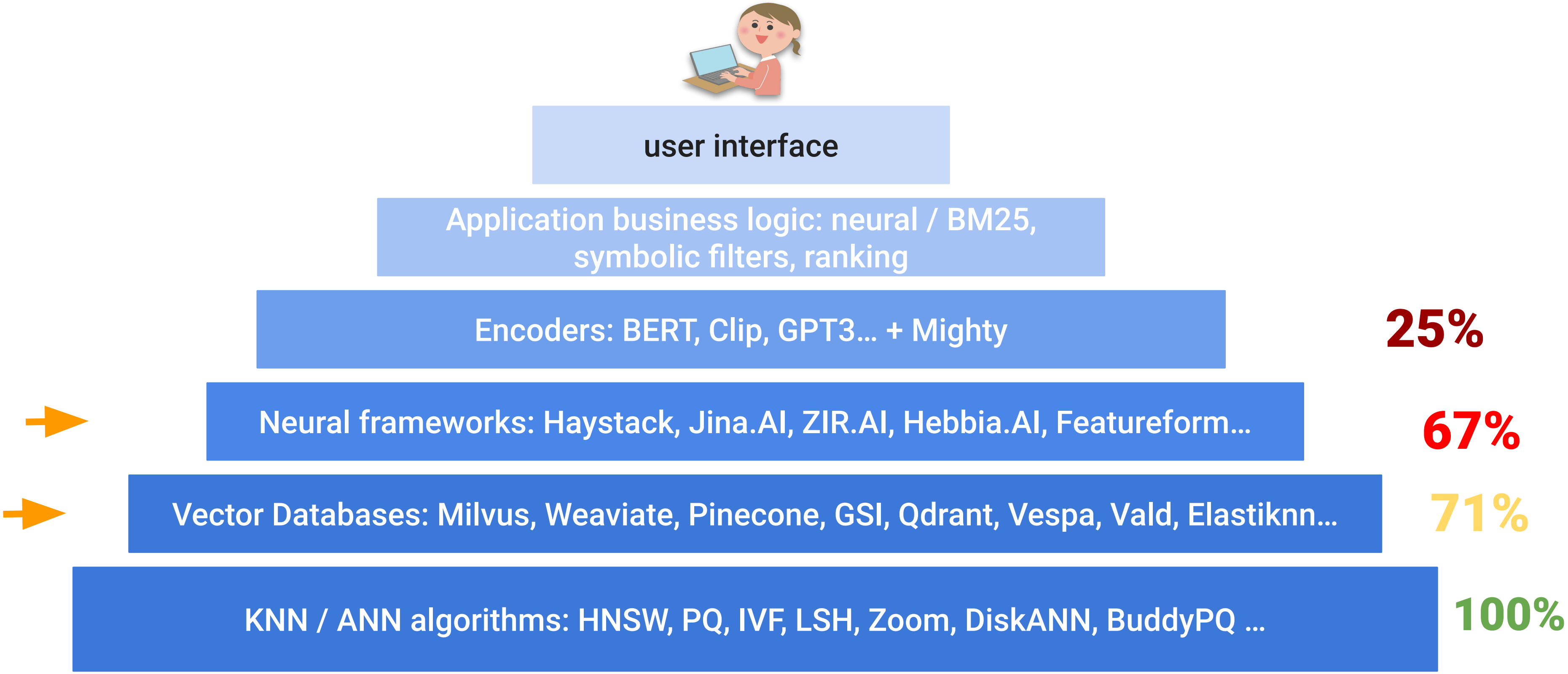
Hybrid Search Methods



Name	Author	Function	Description
CombSUM	Fox and Shaw [15]	$\sum_{d \in D} S(d)$	Score-based — Adds the retrieval scores of documents contained in more than one list and rearranges the order.
CombMNZ	Fox and Shaw [15]	$\lvert d \in D \rvert \cdot \sum_{d \in D} S(d)$	Score-based — Adds the retrieval scores of documents contained in more than one list, and multiplies their sum by the number of lists where the document occurs.
Borda	de Borda [12]	$\frac{n - r(d) + 1}{n}$	Rank-based — Voting algorithm that sums the difference in rank position from the total number of document candidates in each list.
RRF	Cormack et al. [10]	$\sum_{d \in D} \frac{1}{k + r(d)}$	Rank-based — discounts the weight of documents occurring deep in retrieved lists using a reciprocal distribution. The parameter k is typically set to 60.
ISR	Mourao et al. [26]	$\lvert d \in D \rvert \cdot \sum_{d \in D} \frac{1}{r(d)^2}$	Rank-based — inspired by RRF, but discounts documents occurring lower in the ranking more severely.
logISR	Mourao et al. [26]	$\log(\lvert d \in D \rvert) \cdot \sum_{d \in D} \frac{1}{r(d)^2}$	Similar to ISR but with logarithmic document frequency normalization.
RBC	Bailey et al. [3]	$\sum_{d \in D} (1 - \phi)\phi^{r(d)-1}$	Rank-based — discounts the weights of documents following a geometric distribution, inspired by the RBP evaluation metric. [24]

https://rodgerbenham.github.io/bc17-adcs.pdf

Vector Search Pyramid



user interface

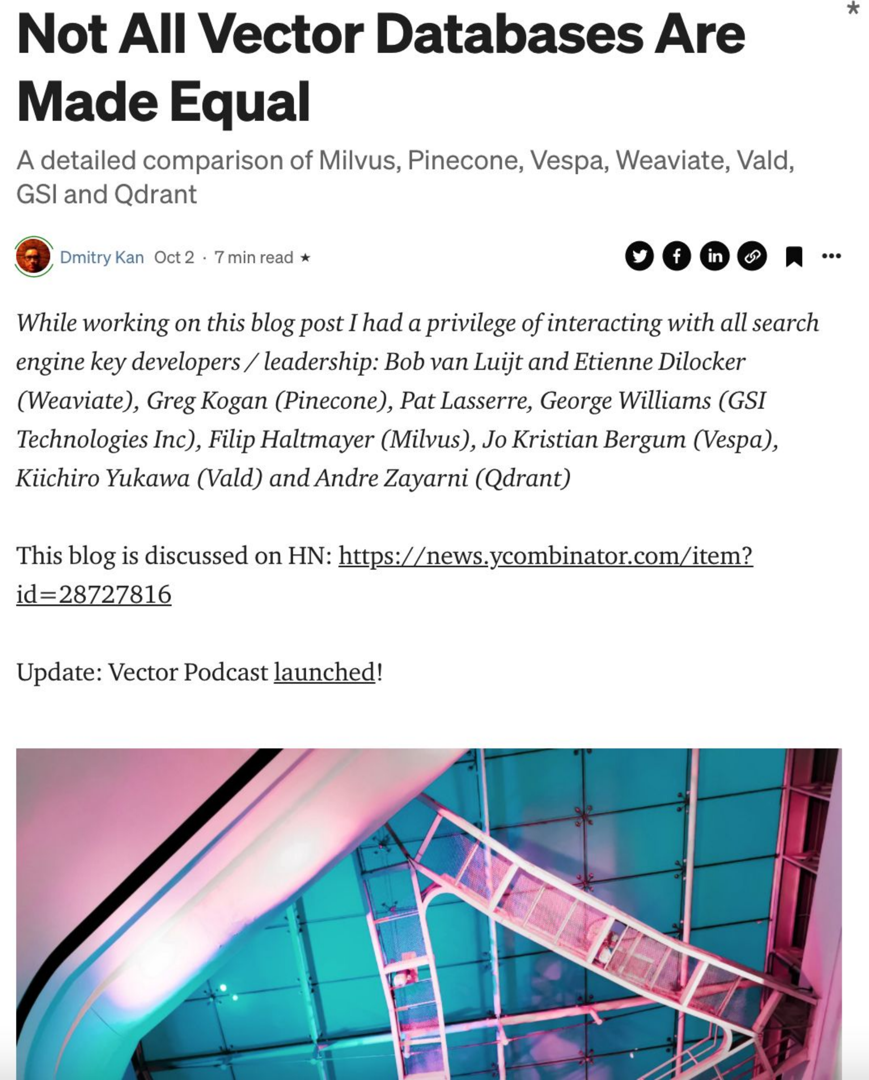
Application business logic: neural / BM25, symbolic filters, ranking

Encoders: BERT, Clip, GPT3... + Mighty 25%

Neural frameworks: Haystack, Jina.AI, ZIR.AI, Hebbia.AI, Featureform... 67%

Vector Databases: Milvus, Weaviate, Pinecone, GSI, Qdrant, Vespa, Vald, Elastiknn... 71%

KNN / ANN algorithms: HNSW, PQ, IVF, LSH, Zoom, DiskANN, BuddyPQ ... 100%



*

Not All Vector Databases Are Made Equal

A detailed comparison of Milvus, Pinecone, Vespa, Weaviate, Vald, GSI and Qdrant

Dmitry Kan Oct 2 · 7 min read ★

While working on this blog post I had a privilege of interacting with all search engine key developers / leadership: Bob van Luijt and Etienne Dilocker (Weaviate), Greg Kogan (Pinecone), Pat Lasserre, George Williams (GSI Technologies Inc), Filip Haltmayer (Milvus), Jo Kristian Bergum (Vespa), Kiichiro Yukawa (Vald) and Andre Zayarni (Qdrant)

This blog is discussed on HN: https://news.ycombinator.com/item?id=28727816

Update: Vector Podcast launched!

Smaller Vector DB players: 71% are Open Source

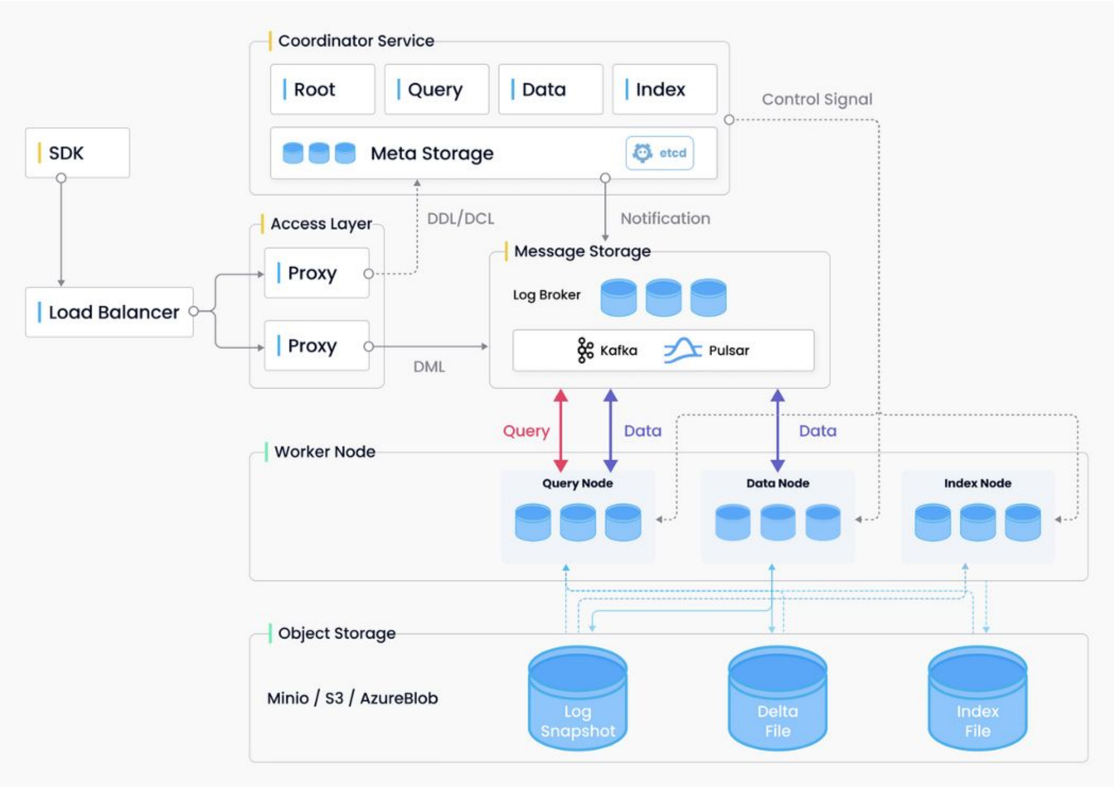
Company	Product	Cloud	Open Source: Y/N	Algorithms
SeMI	Weaviate	Y	Y (Go)	custom HNSW
Pinecone	Pinecone	Y	N	FAISS + own
GSI	APU chip for Elasticsearch / Opensearch	N	N	Neural hashing / Hamming distance
Qdrant	Qdrant	N	Y (Rust)	HNSW (graph)
Yahoo!	Vespa	Y	Y (Java, C++)	HNSW (graph)
Ziliz	Milvus	N	Y (Go, C++, Python)	FAISS, HNSW
Yahoo!	Vald	N	Y (Go)	NGT

Milvus

🌍 milvus.io
💡 self-hosted vector database
🤖 open source

Value proposition:

- attention to scalability of the entire search engine: (re)indexing and search
- ability to index data with multiple ANN algorithms to compare their performance for your use case



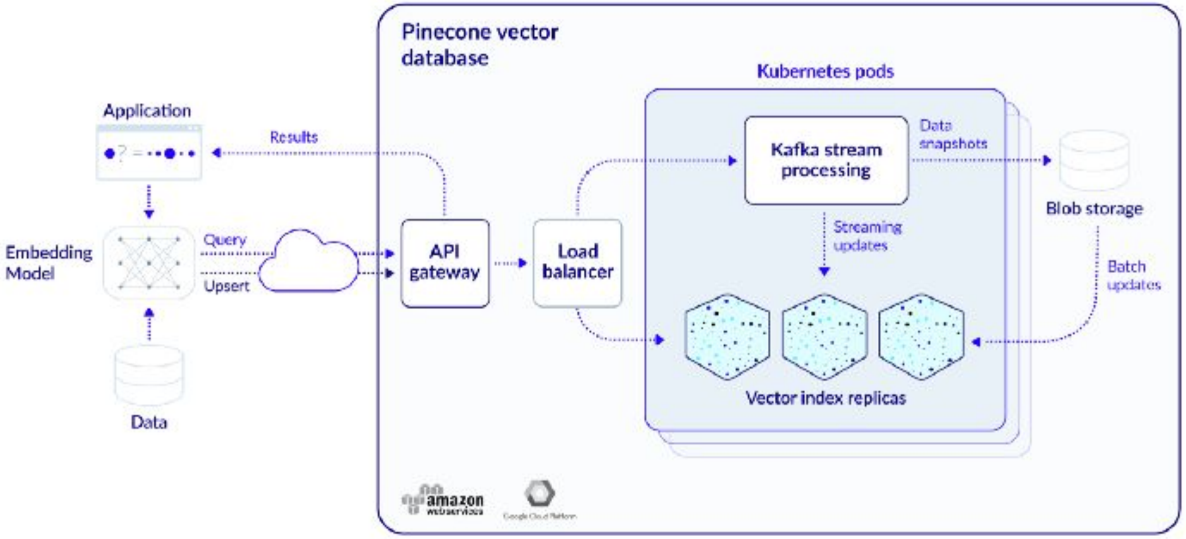
milvus

Pinecone

🌍 pinecone.io
💡 managed vector database
🤖 close source

Value proposition:

- Fully managed vector database
- Single-stage filtering capability: search for your objects (sweaters) + filter by metadata (color, size, price) in one query



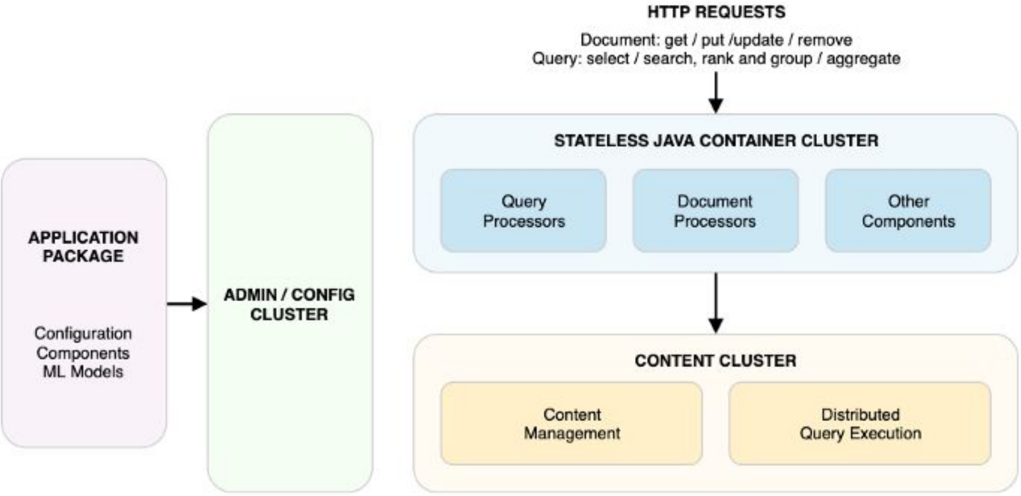
Pinecone

Vespa

🌍 vespa.ai/
💡 managed / self-hosted
🤖 Code: open source

Value proposition:

- low-latency computation over large data sets
- stores and indexes your data so that queries, selection and processing over the data can be performed at serving time
- customizable functionality
- deep data structures geared towards deep-learning like data science, like Tensors



HTTP REQUESTS

Document: get / put /update / remove
Query: select / search, rank and group / aggregate

APPLICATION PACKAGE

Configuration
Components
ML Models

ADMIN / CONFIG CLUSTER

STATELESS JAVA CONTAINER CLUSTER

Query Processors

Document Processors

Other Components

CONTENT CLUSTER

Content Management

Distributed Query Execution

Weaviate

🌍 semi.technology/developers/weaviate/current/

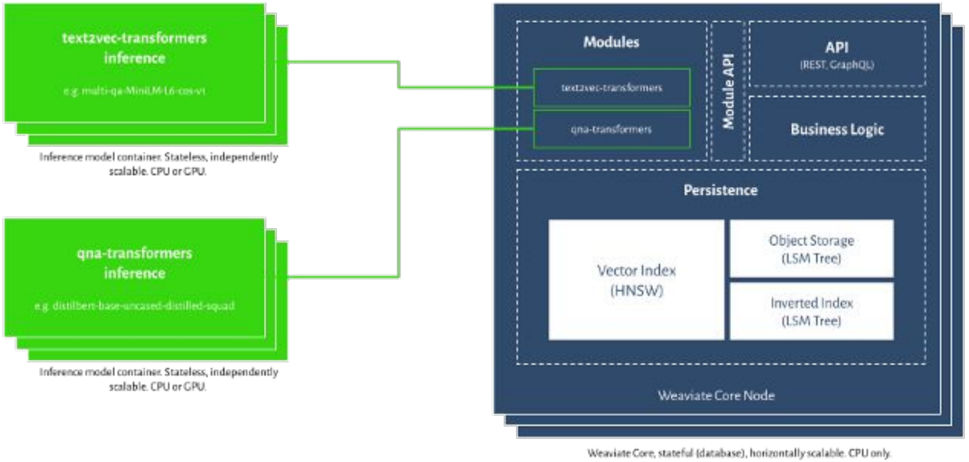
💡 managed / self-hosted

🤖 open source

Value proposition:

- Expressive query syntax
- Graphql-like interface
- combo of vector search, object storage and inverted index
- Wow-effect: Has an impressive question answering component — esp for demos

Weaviate System Level Overview (Example with two modules)


text2vec-transformers inference
e.g. multi-qa-MiniLM-L6-cos-v1
Inference model container. Stateless, independently scalable. CPU or GPU.
qna-transformers inference
e.g. distilbert-base-uncased-distilled-squad
Inference model container. Stateless, independently scalable. CPU or GPU.
Modules
text2vec-transformers
qna-transformers
Module API
API
(REST, GraphQL)
Business Logic
Persistence
Vector Index
(HNSW)
Object Storage
(LSM Tree)
Inverted Index
(LSM Tree)
Weaviate Core Node
Weaviate Core, stateful (database), horizontally scalable. CPU only.

Two modules (text2vec-transformers, qna-transformers) shown as an example. Other modules include vectorization for other media types, entity recognition, spell checking and others.

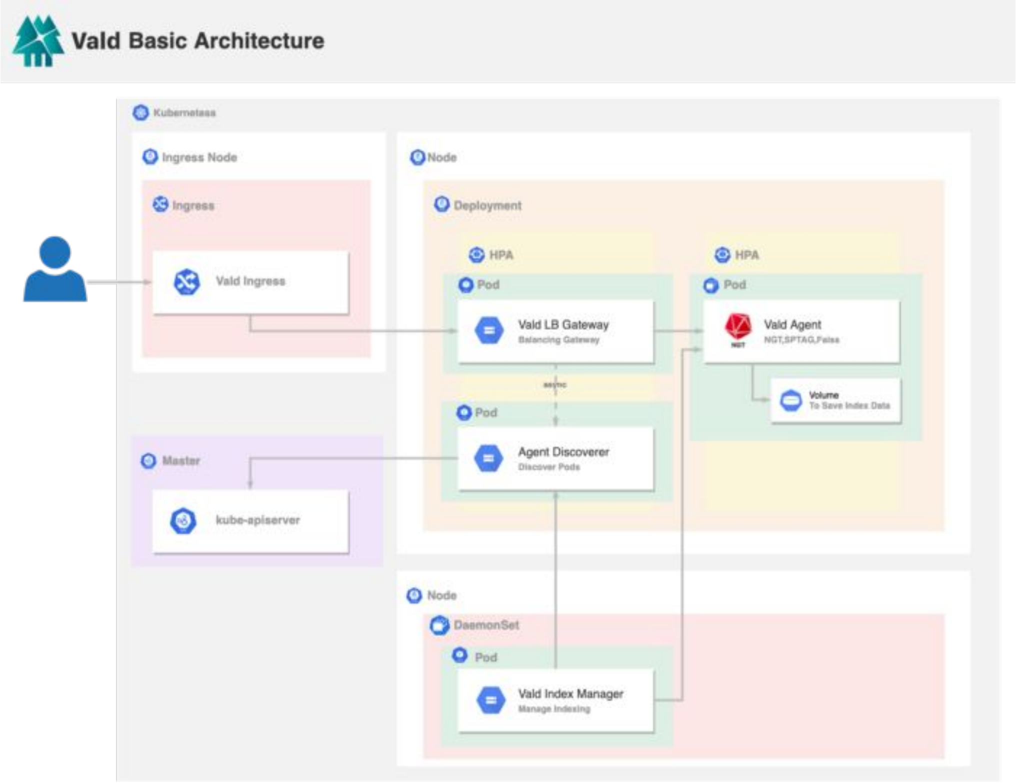
Persistence in Weaviate Core shows one shard as an example. Users can create any number of indices, each index can contain any number of shards. Shards can be distributed and/or replicated across nodes in the cluster. A shard always contains object, inverted and vector storage. Vector storage is not affected by LSM segmentation.

Vald

🌍 Link: vald.vdaas.org/
💡 Type: Self-hosted vector database
🤖 Code: open source

Value proposition:

- Billion-scale
- Cloud-native architecture
- Fastest ANN Algo: NGT
- Custom reranking / filtering algorithm plugins

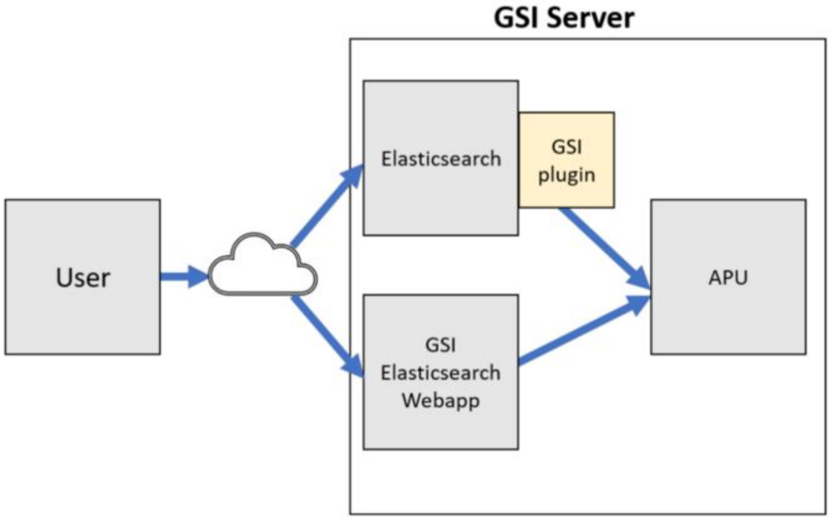


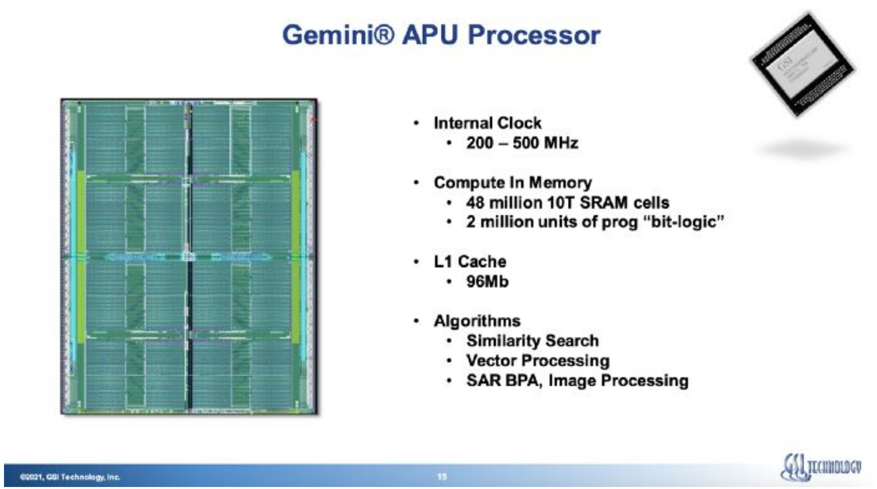
GSI APU

🌎 Link: gsitechnology.com/APU
💡 Type: Vector search hardware backend for your Elasticsearch / OpenSearch
🤖 Code: close source

Value proposition:

- Billion-scale
- Extends your Elasticsearch / OpenSearch capabilities to similarity search
- On-prem / hosted APU board hosted cloud backend





Gemini® APU Processor

- Internal Clock
  - 200 – 500 MHz
- Compute In Memory
  - 48 million 10T SRAM cells
  - 2 million units of prog "bit-logic"
- L1 Cache
  - 96Mb
- Algorithms
  - Similarity Search
  - Vector Processing
  - SAR BPA, Image Processing

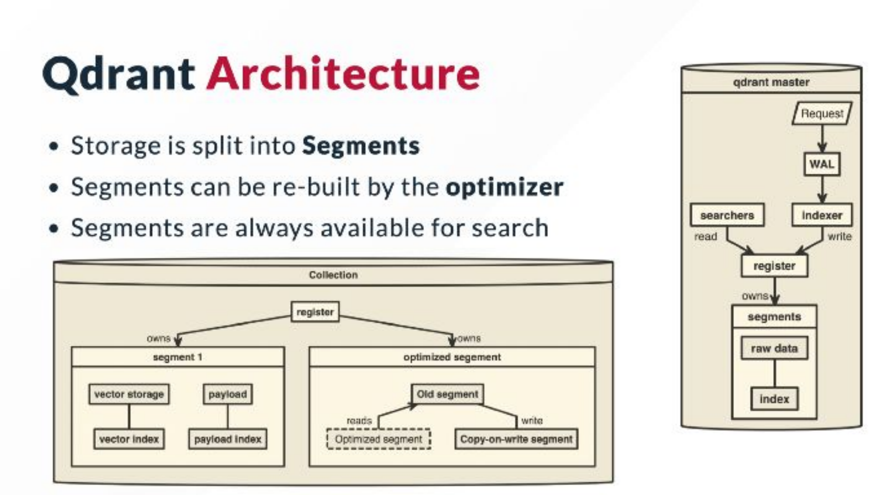
©2021, GSI Technology, Inc. 15

Qdrant

🌍 qdrant.tech/
💡 self-hosted vector database (Cloud in roadmap)
🤖 open source

Value proposition:

- The vector similarity engine with extended filtering support
- dynamic query planning and payload data indexing
- string matching, numerical ranges, geo-locations, and more
- Metric Deep Learning



Qdrant Architecture

- Storage is split into **Segments**
- Segments can be re-built by the **optimizer**
- Segments are always available for search

Semantic search frameworks: 67% Open Source

Company	Product	Open Source: Y/N	Focus
Deepset.ai	Haystack	Y	NLP
Jina.AI	Jina, Hub, Finetuner	Y	NLP, CV, ASR
Featureform	Feature store, EmbeddingHub	Y	All AI verticals
NeuML	txtAI	Y	NLP, CV, ASR
Vector AI	Vector AI	Y	NLP, CV, ASR
ZIR.AI	AI search platform	N	NLP
Hebbia.AI	Knowledge Base	N	NLP
Rasa.ai	Virtual assistants	Y	NLP
Muves.io	Multilingual and multimodal vector search	N	Multilingual search, multimodality

How to pick a vector DB / framework

- Have own engineering?
    - Yes: go for the framework vendor / self-hosted DB
    - No: choose higher-level system, like Hebbia.AI
- Own embedding layer or OK with vector DB doing it?
    - Own: Qdrant, Milvus, Pinecone, GSI, Vespa, Vald
    - In-DB: Weaviate, Vespa
- Heavy focus on NLP?
    - YES: Consider Haystack (deepset)
    - NO: Consider Jina.AI
- Want to quickly test before investing?
    - Yes: ZIR.AI, Hebbia.AI
    - No: Jina.AI, Haystack etc

How to pick a vector DB / framework

- Want to HOST or fine with MANAGED?
  - HOST: Vespa, Vald, Milvus, Qdrant
  - MANAGED: Pinecone, Weaviate, GSI, Hebbia.AI, ZIR.AI

Use cases

Specific:

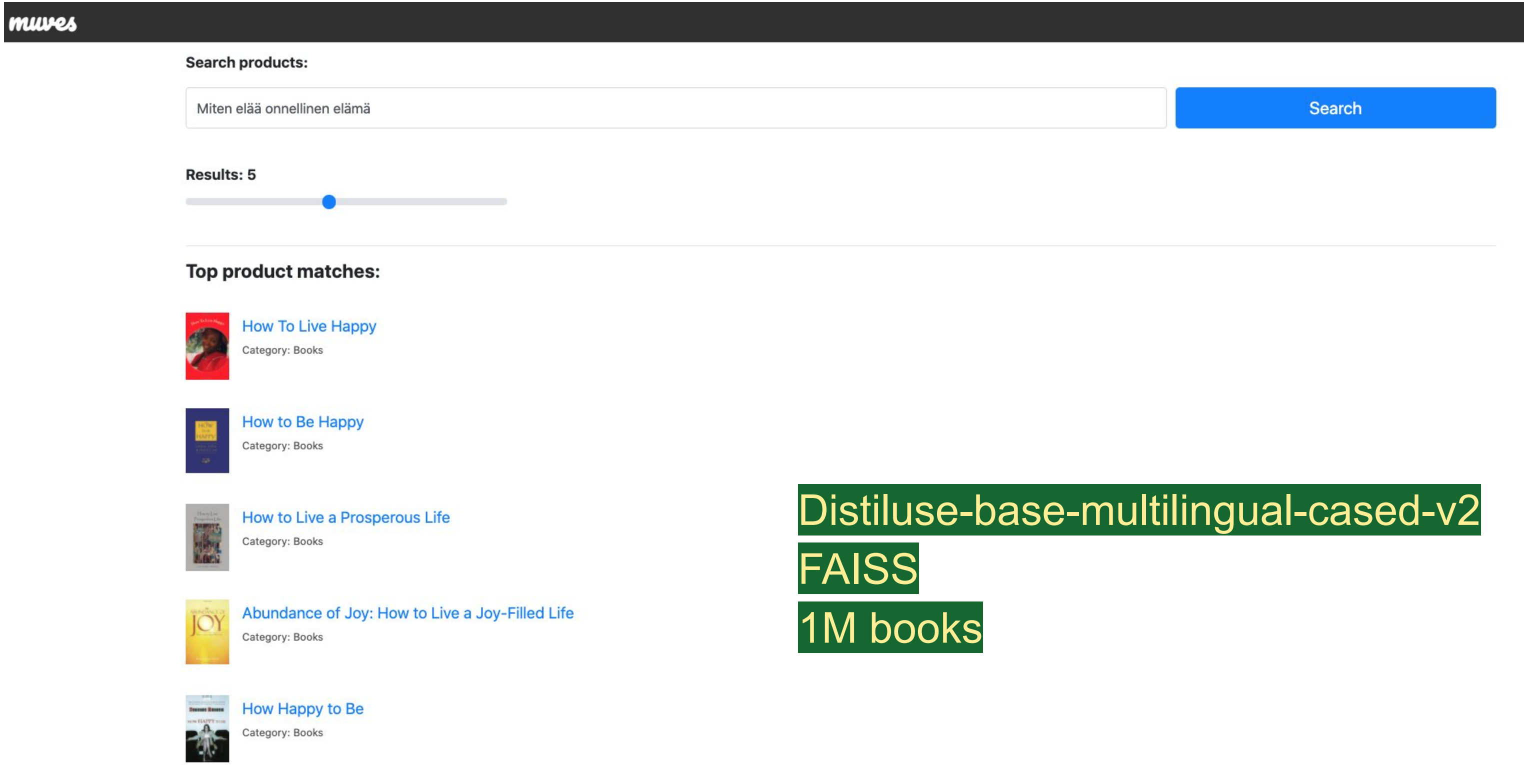
- Image similarity (CLIP retrieval with KNN search)
- Multilingual search (Muves)
- Question answering (Techcrunch Weaviate demo, Deepset)
- Recommenders: Facebook news feed
- Google Talk to Books
- Car image search (Classic.com)
- E-commerce: multimodal search (Muves+GSI APU)

Broad:

- Metric learning: use case for Vector DBs
- Semantic search: *ride-sharing*
- Anomaly detection
- Classification
- Multi-stage ranking

Demo: Muves books search

muves.io



Search products:

Miten elää onnellinen elämä

Results: 5

Top product matches:

How To Live Happy
Category: Books

How to Be Happy
Category: Books

How to Live a Prosperous Life
Category: Books

Abundance of Joy: How to Live a Joy-Filled Life
Category: Books

How Happy to Be
Category: Books

Distiluse-base-multilingual-cased-v2

FAISS

1M books

For the demo we used a multilingual CLIP from Huggingface and a 10M subset from LAION-400M dataset

Hugging Face

Multilingual CLIP model for image search

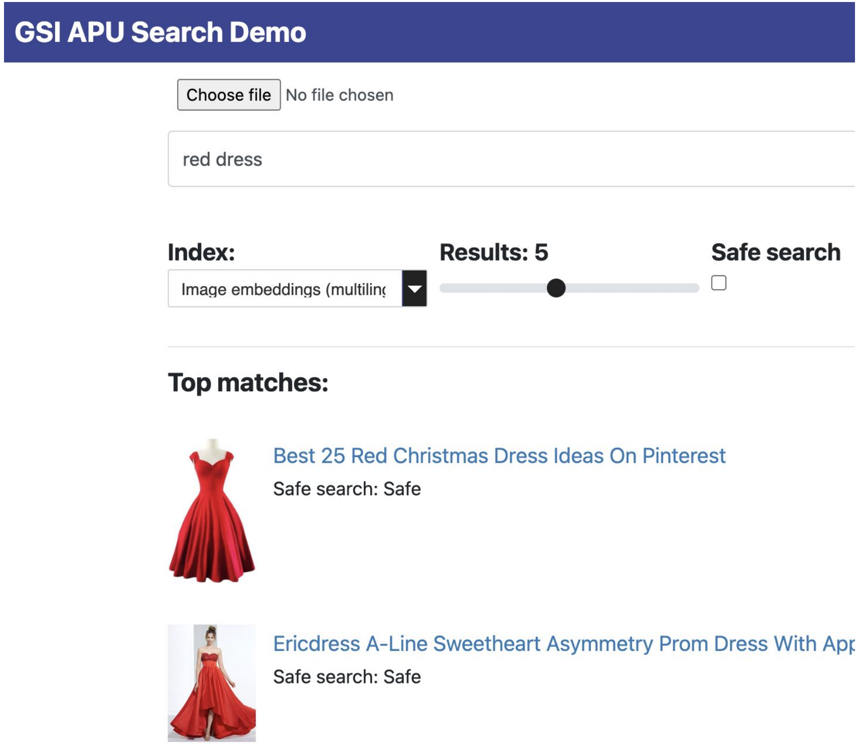
Multilingual Sentence Transformers for text embeddings

LAION-400M

The world's largest openly available image-text-pair dataset with 400 million samples.

Demo: Muves multilingual & multimodal search

muves.io



GSI APU Search Demo

Choose file No file chosen

red dress

Index: Image embeddings (multilinc

Results: 5

Safe search □

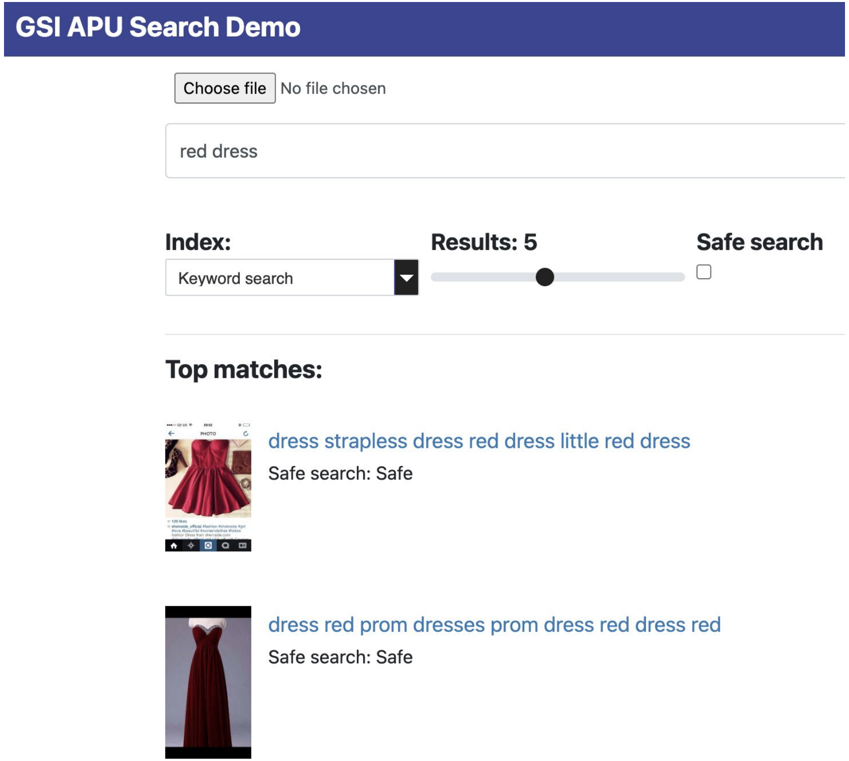
Top matches:

Best 25 Red Christmas Dress Ideas On Pinterest

Safe search: Safe

Ericdress A-Line Sweetheart Asymmetry Prom Dress With App

Safe search: Safe



GSI APU Search Demo

Choose file No file chosen

red dress

Index: Keyword search

Results: 5

Safe search □

Top matches:

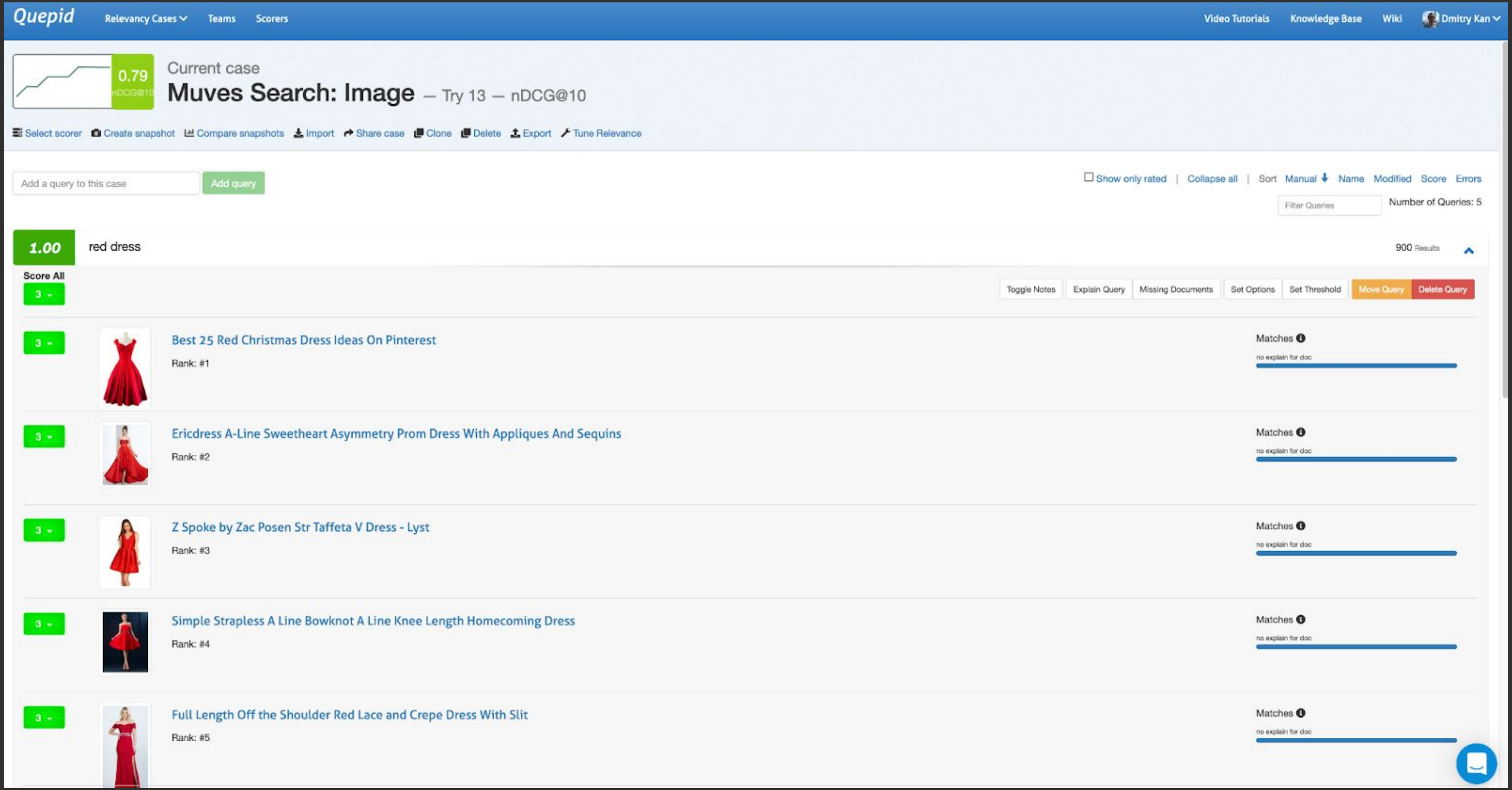
dress strapless dress red dress little red dress

Safe search: Safe

dress red prom dresses prom dress red dress red

Safe search: Safe

Image embedding search: NDCG@10



Current case

Muves Search: Image — Try 13 — nDCG@10

Number of Queries: 5

1.00 red dress — 900 Results

Score All: 3

Score	Title	Rank	
3	Best 25 Red Christmas Dress Ideas On Pinterest	Rank: #1	Matches no explain for doc
3	Ericdress A-Line Sweetheart Asymmetry Prom Dress With Appliques And Sequins	Rank: #2	Matches no explain for doc
3	Z Spoke by Zac Posen Str Taffeta V Dress - Lyst	Rank: #3	Matches no explain for doc
3	Simple Strapless A Line Bowknot A Line Knee Length Homecoming Dress	Rank: #4	Matches no explain for doc
3	Full Length Off the Shoulder Red Lace and Crepe Dress With Slit	Rank: #5	Matches no explain for doc

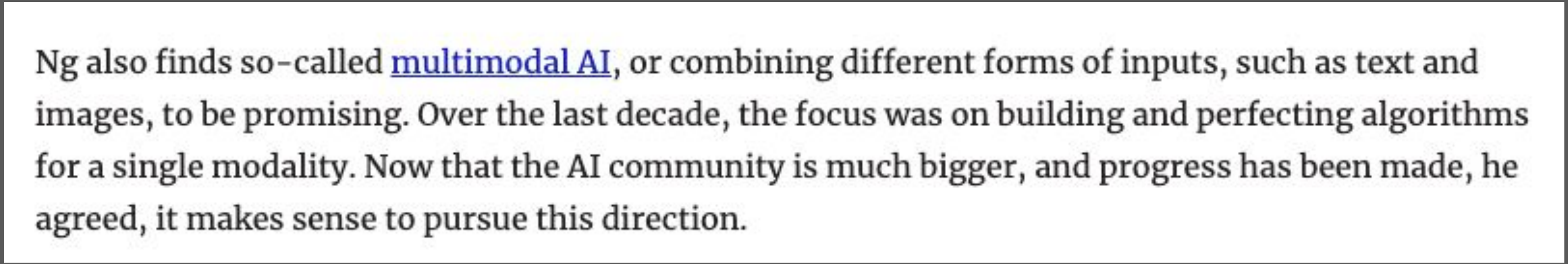
Where things are going

- BM25 / Vector search: how to combine?
  - Vespa’s talk on BBUZZ’22
  - Jina demos with text/image blending
- Efficient embeddings
  - Mighty
  - GPU vs CPU (cost)
  - Latency at scale
- Going multimodal
  - Muves: multimodal search + hardware acceleration
  - Search inside audio and blend with text-based matching
  - Search complex objects, like 3D mesh, polygon similarity
- Model fine-tuning / selection
  - BEIR paper
  - Play with Jina Finetuner
  - Pre-trained models on your domain, like CLIP
  - Large LMs, becoming practical, like Atlas
- Choose strategy for your vector search
  - Add new vector DB or enable a dense_vector in Elasticsearch / OpenSearch / Solr
  - Doc2query – no vector search is involved!
  - Precise vs Explorative search
  - Use cases in your product
  - MLOps

Trends in ML at large

- Model hubs (e.g. Hugging Face) → ML community shares progress quickly (similar to what GitHub did to sharing code)
- Deep Learning → multimodal: CLIP (text from images), DALL-E (images from text)
- MLOps optimize experimentation and deployments: determined.ai, DVC, MLflow / Kubeflow
- Big Language Models

Multimodal search



Ng also finds so-called multimodal AI, or combining different forms of inputs, such as text and images, to be promising. Over the last decade, the focus was on building and perfecting algorithms for a single modality. Now that the AI community is much bigger, and progress has been made, he agreed, it makes sense to pursue this direction.

Get practical

- Code: https://github.com/DmitryKey/bert-solr-search
- Supported Engines: Solr, Elasticsearch, OpenSearch, GSI, [hnswlib]
- Supported LMs: BERT, SBERT, [theoretically any]

Thank you! 🧡

twitter.com/DmitryKan

youtube.com/c/VectorPodcast

https://spoti.fi/3sRXcdn

Links

1. Martin Fowler’s talk on NoSQL databases: https://www.youtube.com/watch?v=qI_g07C_Q5I
2. Neural search vs Inverted search https://trends.google.com/trends/explore?date=all&q=neural%20search,inverted%20search
3. Not All Vector Databases Are Made Equal https://towardsdatascience.com/milvus-pinecone-vespa-weaviate-vald-gsi-what-unites-these-buzz-words-and-what-makes-each-9c65a3bd0696
4. HN thread: https://news.ycombinator.com/item?id=28727816
5. A survey of PQ and related methods: https://faiss.ai/
6. Vector Podcast on YouTube: https://www.youtube.com/c/VectorPodcast
7. Vector Podcast on Spotify: https://open.spotify.com/show/13JO3vhMf7nAqcpvlIgOY6
8. Vector Podcast on Apple Podcasts: https://podcasts.apple.com/us/podcast/vector-podcast/id1587568733
9. BERT, Solr, Elasticsearch, OpenSearch, HNSWlib – in Python: https://github.com/DmitryKey/bert-solr-search
10. Speeding up BERT search in Elasticsearch: https://towardsdatascience.com/speeding-up-bert-search-in-elasticsearch-750f1f34f455

Links

1. Merging Knowledge Graphs with Language Models: https://www.microsoft.com/en-us/research/publication/jaket-joint-pre-training-of-knowledge-graph-and-language-understanding/
2. Fusing Knowledge into Language Models: https://cs.stanford.edu/people/cgzhu/ in particular: https://drive.google.com/file/d/1-jvKce5rcBp_DdkPv8ijLeFW5N2XrapS/view
3. Players in Vector Search: Algorithms, Software and Use Cases https://dmitry-kan.medium.com/players-in-vector-search-video-2fd390d00d6
4. How to Choose a Vector Database: https://dmitry-kan.medium.com/how-to-choose-a-vector-database-8c6e6f0f8f8b
5. Zoom paper: https://arxiv.org/abs/1809.04067
6. Risk-Reward Trade-offs in Rank Fusion: https://rodgerbenham.github.io/bc17-adcs.pdf