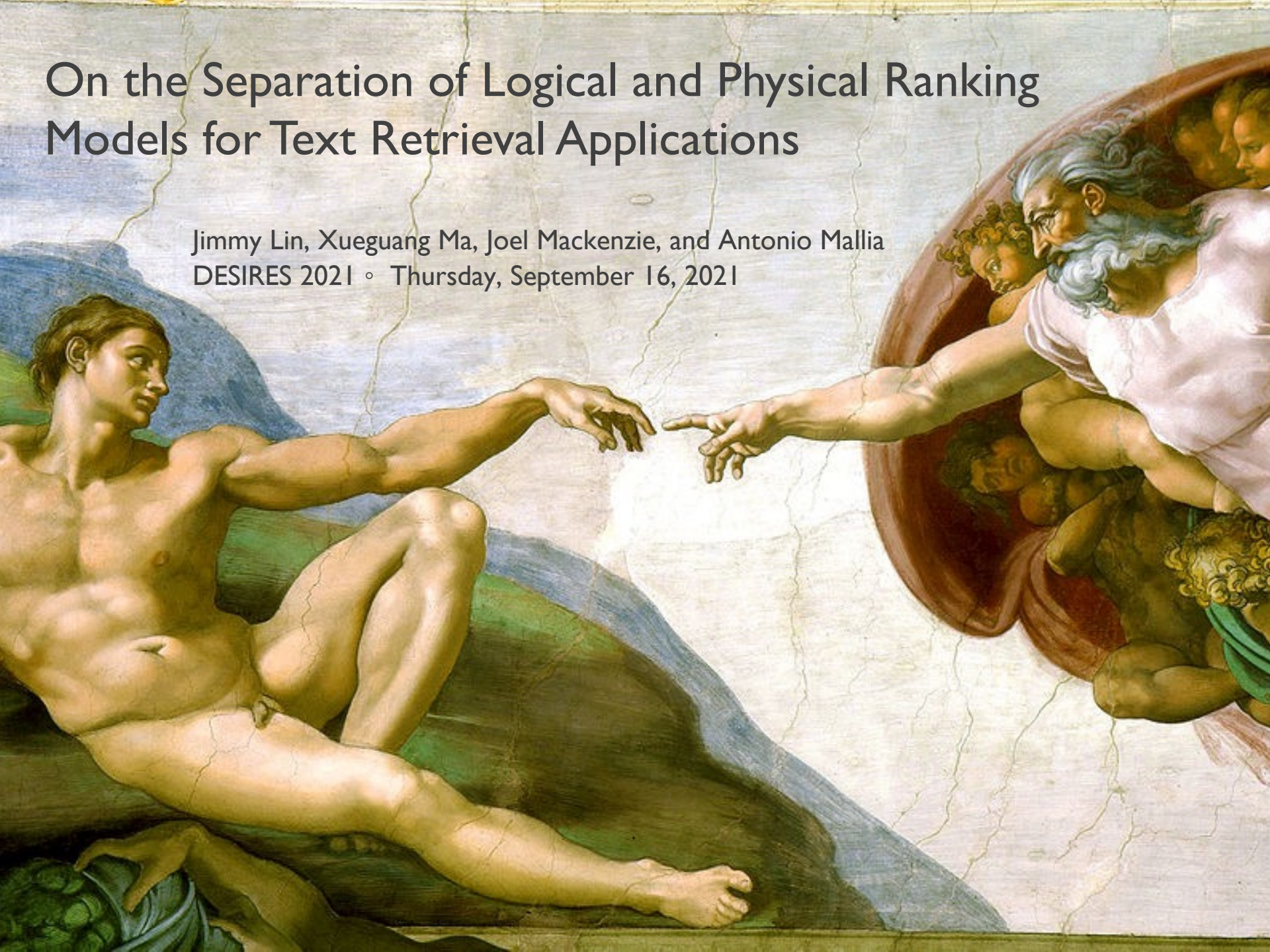
On the Separation of Logical and Physical Ranking Models for Text Retrieval Applications

Jimmy Lin, Xueguang Ma, Joel Mackenzie, and Antonio Mallia

DESIRES 2021 ◦ Thursday, September 16, 2021



What's the opposite of logical?

illogical?

Wrong!

physical, of course!

tl;dr –

Information Retrieval breaks down into two components:

The *Logical* Scoring Model

(how to compute query-document scores)

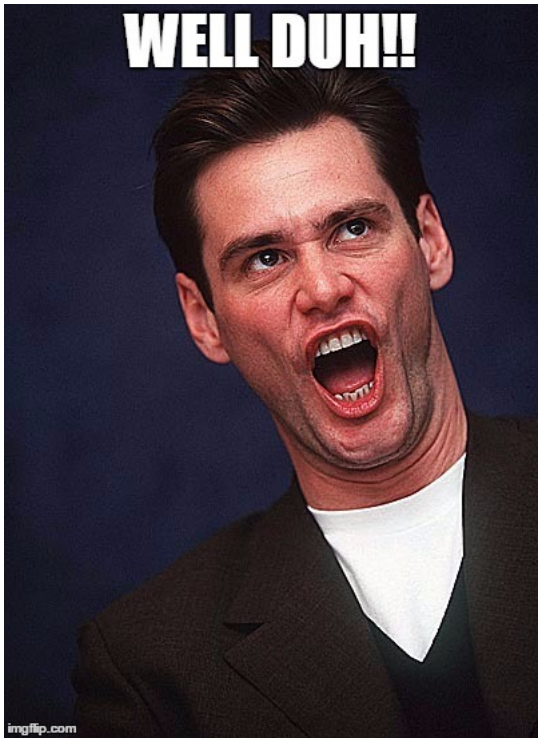
$$s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$$

The *Physical* Retrieval Model

(how to retrieve top-*k* scoring documents from corpus)

$$\arg \underset{d \in \mathcal{D}}{\text{top-k}}\ \phi(\eta_q(q), \eta_d(d))$$

(of course, borrowing from database systems)



Goal: to convince you that this isn't just pointless symbol manipulation.

Source: flickr (tapasinthesun/49114923568)

Information Retrieval in Two Steps

document (*ad hoc*) retrieval

Retrieve

Rerank

Information Retrieval in Two Steps

Retrieve

Learning to Rank

Multi-stage ranking
Early exits
Selective evaluation
…

Variations on a theme!

Source: Wikipedia

Information Retrieval in Two Steps

Retrieve

Learning to Rank

Information Retrieval in Two Steps

Retrieve

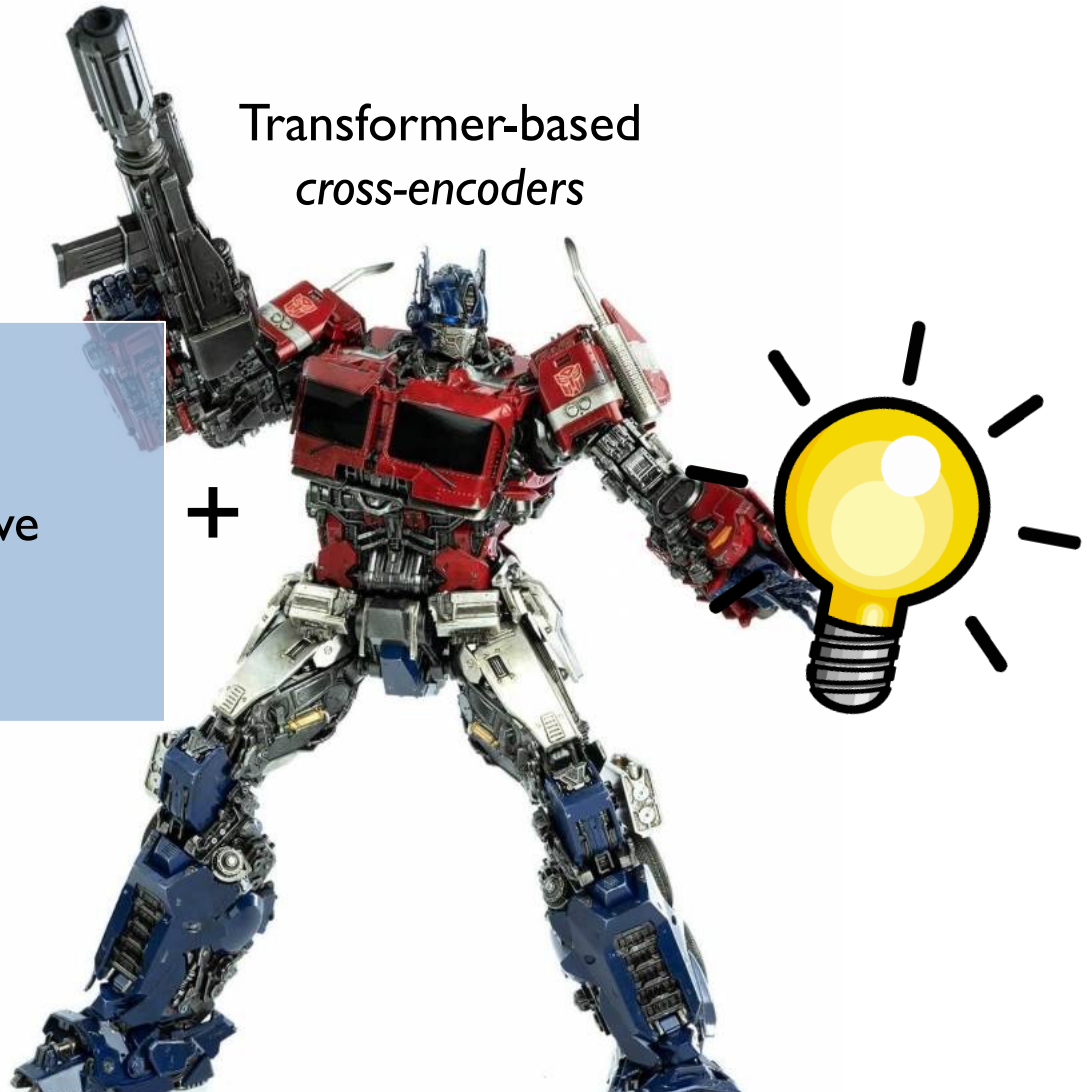
+

Information Retrieval in Two Steps

Retrieve

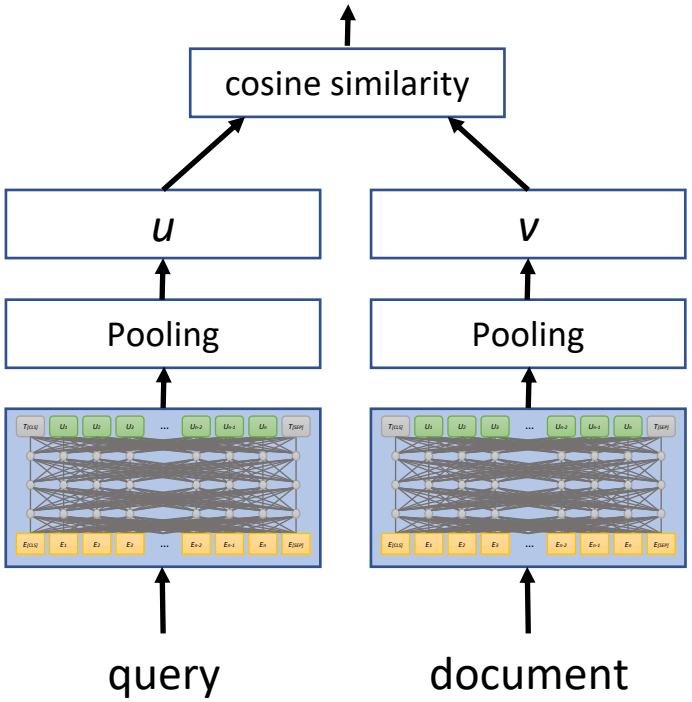
+

Transformer-based
cross-encoders



Source: Wikipedia

Enter dense retrieval…



Why is dense retrieval exciting?

Different ways of using transformers
(bi-encoders vs. cross-encoders)

Interesting effectiveness-efficiency tradeoffs
(less effective than cross-encoders, but much faster)

Versatile
(single-stage ranking and first-stage ranking)

Different "software stack"
(inverted indexes vs. HNSW)

What's the relationship between dense retrieval and sparse retrieval?

Hint: I've already shared the answer!

Information Retrieval breaks down into two components:

The *Logical* Scoring Model

(how to compute query-document scores)

$$s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$$

The *Physical* Retrieval Model

(how to retrieve top-*k* scoring documents from corpus)

$$\arg \underset{d \in \mathcal{D}}{\text{top-k}}\ \phi(\eta_q(q), \eta_d(d))$$

BTW, this isn't a new idea…

Previous Work

Discussion of representational separation in IR

Fuhr. Models for integrated information retrieval and database systems. 1996.

Implementing retrieval directly in databases

Héman et al. Efficient and flexible information retrieval using MonetDB/X100. *CIDR 2007*.

GeeseDB!

Let's apply this to analyze dense and sparse retrieval…

The Logical Scoring Model

$$s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$$

The Physical Retrieval Model

$$\underset{d \in \mathcal{D}}{\text{arg top-k}}\ \phi(\eta_q(q), \eta_d(d))$$

	BM25 (sparse)	DPR (dense)
	BoW representations	transformer-based representations
	inner-product	inner-product
	Inverted index + DaaT query evaluation (e.g., Lucene)	HNSW (e.g., Faiss)

"Traditional" tight coupling

Why?

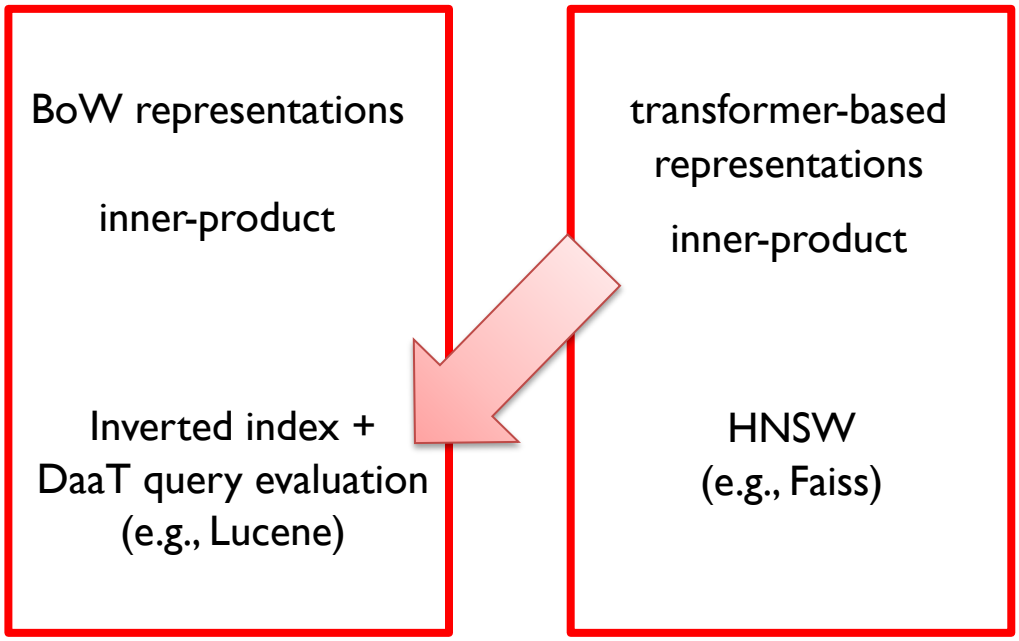
Other combinations are possible!

The Logical Scoring Model

$$s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$$

The Physical Retrieval Model

$$\underset{d \in \mathcal{D}}{\text{arg top-k}}\ \phi(\eta_q(q), \eta_d(d))$$



Teofili and Lin. Lucene for Approximate Nearest-Neighbors Search on Arbitrary Dense Vectors. *arXiv:1910.10208*, 2019.

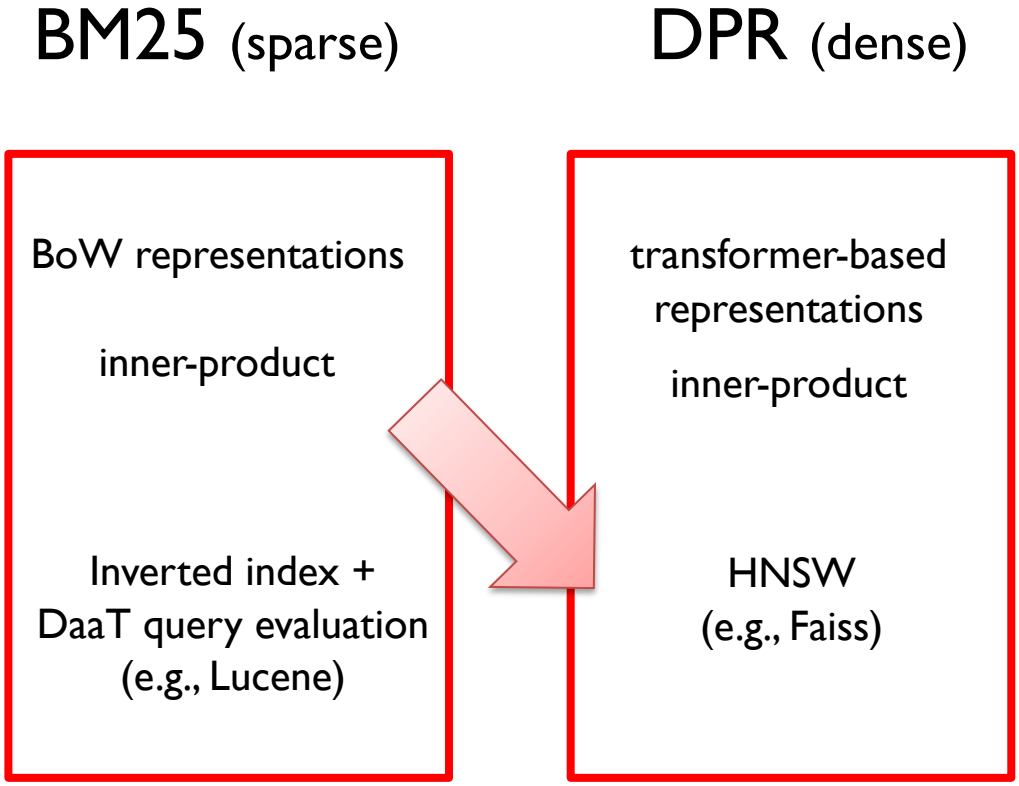
Other combinations are possible!

The Logical Scoring Model

$$s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$$

The Physical Retrieval Model

$$\arg\ \text{top-k}_{d \in \mathcal{D}}\ \phi(\eta_q(q), \eta_d(d))$$



Tu et al. Approximate Nearest Neighbor Search and Lightweight Dense Vector Reranking in Multi-Stage Retrieval Architectures. *ICTIR 2020*.

Method	Quality	Time	Space
	MRR@10	Latency (ms)	Index Size (MB)
Anserini (Lucene)			
(1a) Bag of words	0.187	40.1	661
(1b) doc2query–T5	0.277	62.8	1036
(1c) DeepImpact (quantized)	0.325	244.1	1417
PISA			
(2a) Bag of words	0.187	8.3	739
(2b) doc2query–T5	0.276	11.9	1150
(2c) DeepImpact (quantized)	0.326	19.4	1564
nmslib HNSW			
(3a) DeepImpact	0.299	21.9	6686
(3b) DeepImpact (quantized)	0.298	22.5	6686

Table 1
Experimental results on the development queries of the MS MARCO passage ranking test collection.

Same logical scoring model, different physical retrieval models!

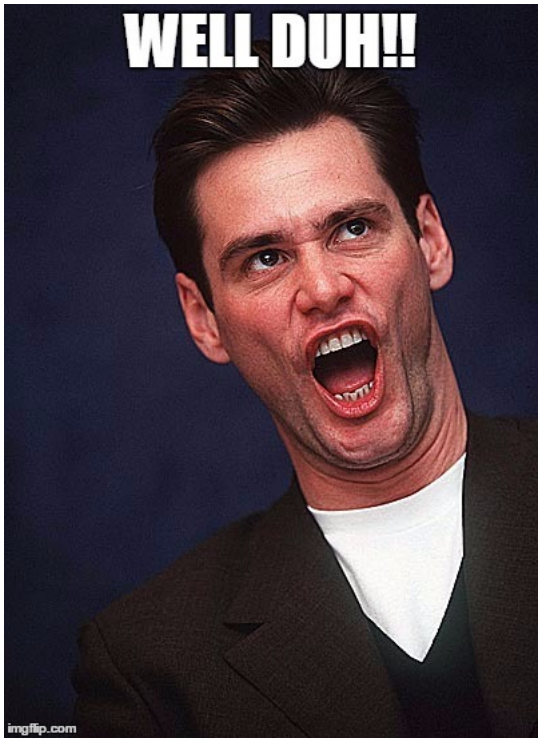
Different quality-time-space tradeoffs!

Mallia et al. Learning Passage Impacts for Inverted Indexes. *SIGIR 2021*.

Method	Quality	Time	Space
	MRR@10	Latency (ms)	Index Size (MB)
Anserini (Lucene)			
(1a) Bag of words	0.187	40.1	661
(1b) doc2query–T5	0.277	62.8	1036
(1c) DeepImpact (quantized)	0.325	244.1	1417
PISA			
(2a) Bag of words	0.187	8.3	739
(2b) doc2query–T5	0.276	11.9	1150
(2c) DeepImpact (quantized)	0.326	19.4	1564
nmslib HNSW			
(3a) DeepImpact	0.299	21.9	6686
(3b) DeepImpact (quantized)	0.298	22.5	6686

Table 1
Experimental results on the development queries of the MS MARCO passage ranking test collection.

So? PISA dominates in all tradeoffs!



My goal: to convince you that this isn't just pointless symbol manipulation.

tl;dr –

Information Retrieval breaks down into two components:

The *Logical* Scoring Model

(how to compute query-document scores)

$$s(q, d) \triangleq \phi(\eta_q(q), \eta_d(d))$$

The *Physical* Retrieval Model

(how to retrieve top-*k* scoring documents from corpus)

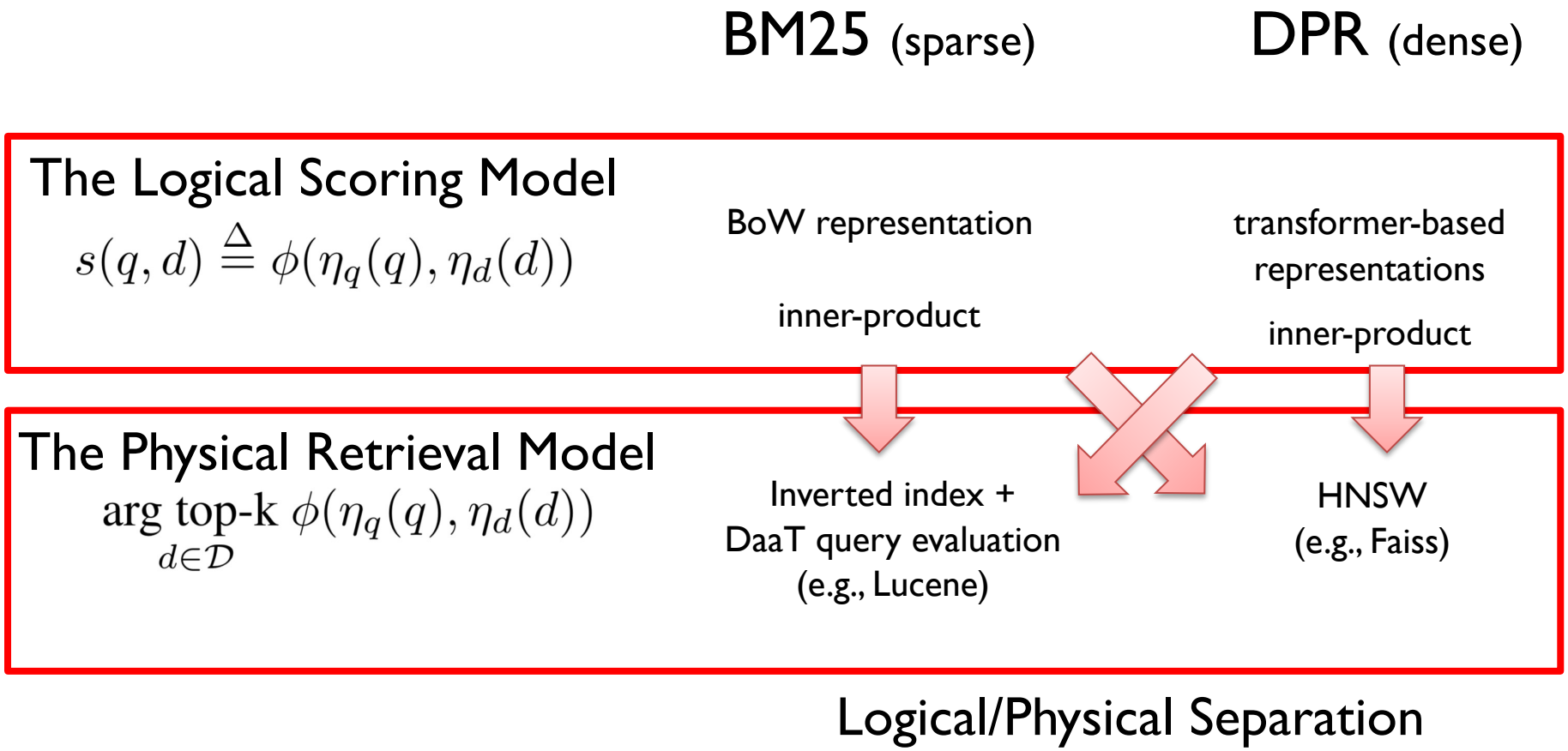
$$\arg \underset{d \in \mathcal{D}}{\text{top-k}}\ \phi(\eta_q(q), \eta_d(d))$$

This provides a nice conceptual framework to think about dense/sparse retrieval!

	BM25 (sparse)	DPR (dense)
The Logical Scoring Model $s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$	BoW representation inner-product	transformer-based representations inner-product
The Physical Retrieval Model $\underset{d \in \mathcal{D}}{\arg \text{top-k}}\ \phi(\eta_q(q), \eta_d(d))$	Inverted index + DaaT query evaluation (e.g., Lucene)	HNSW (e.g., Faiss)

	BM25 (sparse)	DPR (dense)
The Logical Scoring Model $s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$	BoW representation inner-product	transformer-based representations inner-product
The Physical Retrieval Model $\arg\ \text{top-k}_{d \in \mathcal{D}}\ \phi(\eta_q(q), \eta_d(d))$	Inverted index + DaaT query evaluation (e.g., Lucene)	HNSW (e.g., Faiss)

"Traditional" tight coupling


BM25 (sparse)
DPR (dense)
The Logical Scoring Model
$s(q,d) \triangleq \phi(\eta_q(q), \eta_d(d))$
BoW representation
transformer-based representations
inner-product
inner-product
The Physical Retrieval Model
$\arg \text{top-k}_{d \in \mathcal{D}} \phi(\eta_q(q), \eta_d(d))$
Inverted index + DaaT query evaluation (e.g., Lucene)
HNSW (e.g., Faiss)
Logical/Physical Separation

Source: flickr (tapasinthesun/49114923568)

Dawn of a new era?

It's an exciting time to do research!

Questions?