# Question Answering

Lecture 12

Announcement

- Plan for the reminder of the course
  - Week 12 (today): QA + RAG
  - Week 13: RAG + Agent basics
  - Week 14: Review
- Homework:
  - HW4 release today --- regular homework
  - HW5: problem set
    - Gives you a feel of the final exam
  - ~~HW6~~: max(HW)

Modern QA from text

The common person's view? [From a novel]
"I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota … I'm the first guy to get the modem humming. But **as a source of information, it sucks**. You got **a billion pieces of data**, struggling to be heard and seen and downloaded, and **anything I want to know seems to get trampled underfoot in the crowd**."
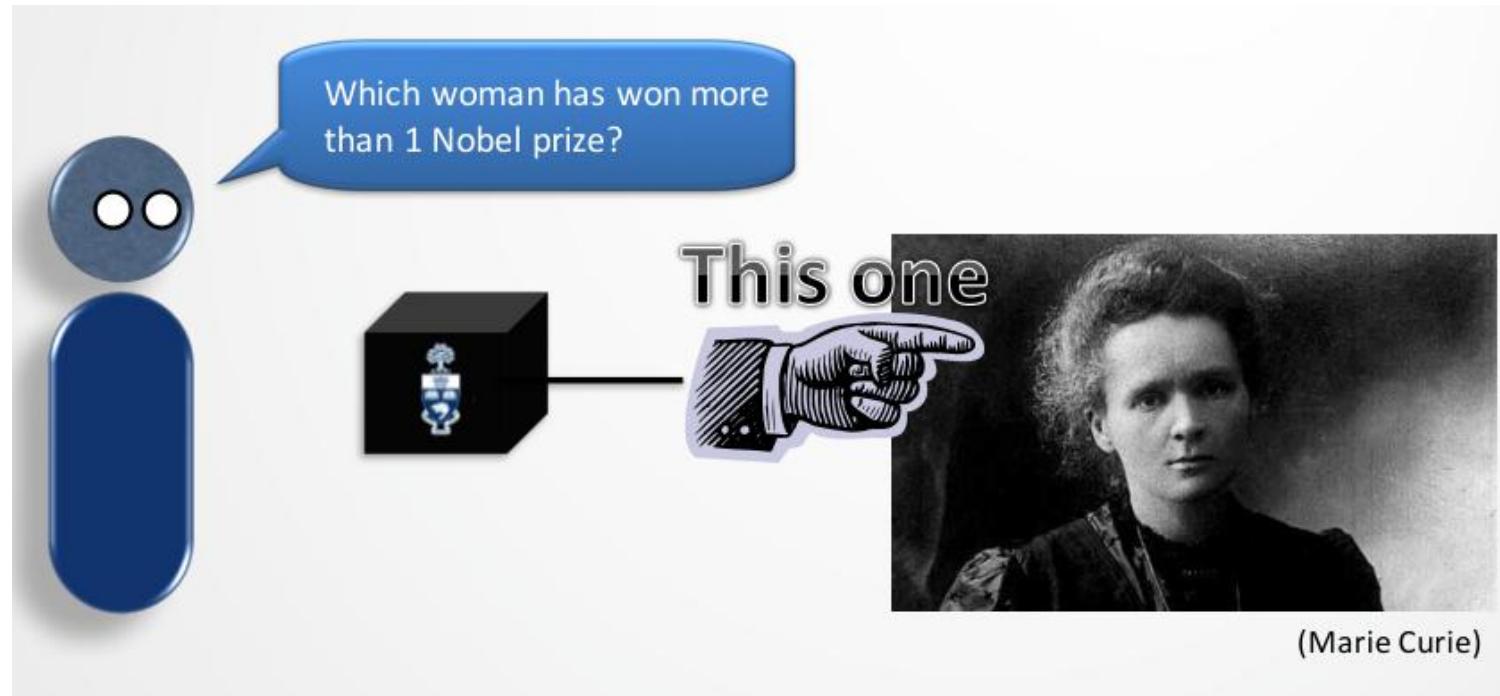
M. Marshall. The Straw Men. HarperCollins Publishers, 2002.

- An idea originating from the IR community.
- With massive collections of full-text documents, simply finding **relevant documents** is of limited use: we want **answers** from textbases.
- QA: give the user a (short) answer to their question, perhaps supported by evidence.

Outline

- Intro to QA
- QA & IR before deep learning
- QA & IR with deep learning (BERT, dense retrieval…)
- RAG: QA with LLM

Question Answering (QA)

(Marie Curie)

- Question Answering (QA) usually involves a specific answer to a question.

Information Retrieval (IR) and QA

One strategy is to turn QA into information retrieval (IR) and let the human complete the task.

6

Question Answering (QA)

Knowledge-based QA

1. Build a structured semantic representation of the query.
   - Extract times, dates, locations, entities using regular expressions.
   - Fit to well-known templates.
2. Query databases with these semantics.
   - Ontologies (Wikipedia infoboxes).
   - Restaurant review databases.
   - Calendars.
   - ...

IR-based QA

where is the louvre museum located

To exit full screen, press and hold `Esc`

Sign in

All | Images | Maps | Videos | News | Shopping | Short videos | More

Tools

Results for **Paris, France** · Choose area

75001 Paris, France

Louvre Museum, Address

Wikipedia
https://en.wikipedia.org › wiki › Louvre

Louvre

The Louvre museum is located **inside the Louvre Palace, in the center of Paris**, adjacent to the Tuileries Gardens. The two nearest Métro stations are Louvre ...

Louvre Palace · Louvre Pyramid · Louvre Abu Dhabi · Art museum

People also ask

Where is the Louvre located exactly?

How far apart are the Louvre and Eiffel Towers?

Is the Louvre where the Mona Lisa is?

How much does it cost to get into the Louvre?

Feedback

Le Louvre
https://www.louvre.fr › visit › map-entrances-directions
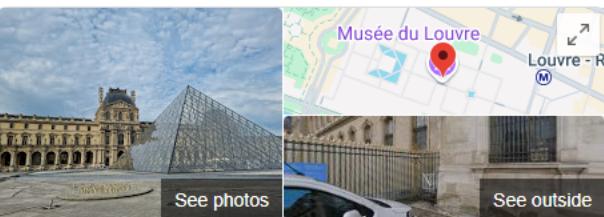
Map, entrances & directions - - All roads lead to the Louvre

An underground car park is located at **1 Avenue du Général Lemonier**, from which you can access the museum via the Galerie du Carrousel entrance.

Britannica
https://www.britannica.com › Visual Arts › Painting

Louvre | History, Collections, & Facts

See photos | See outside

Musée du Louvre

Louvre - R

Louvre Museum

Website | Directions | Save

4.7 ★★★★★ 325,915 Google reviews

Museum in Paris, France

SEE TICKETS

Sponsored

Withlocals

Louvre at Night: Explore with a Local

$125 · 5.0 ★ (1.4K)

The Louvre, or the Louvre Museum, is a national art museum in Paris, France, and one of the most famous museums in the world. Wikipedia

Departments: Librairie-Boutique du Musée du Louvre

Address: 75001 Paris, France

Founded: August 10, 1793

Hours: Closed · Opens 9 a.m. · More hours

Director: Laurence des Cars

Visitors: 8.9 million (2023): Ranked 1st nationally; Ranked 1st globally

Phone: +33 1 40 20 53 17

Subsidiary: Louvre Conservation Center

Curator: Marie-Laure de Rochebrune

Suggest an edit

10

All · Images · Videos · Shopping · News · Forums · Web · More · Tools

Sign in

Louvre Museum / Artworks

Mona Lisa	Venus de Milo	Winged Victory of Samothrace	Liberty Leading the People	Psyche Revived by Cupid's Kiss	The Raft of the Medusa	The Coronation of Napoleon	The Wedding at Cana	The Seated Scribe	The Virgin of the Rocks	La Belle Ferronnière	Oath of the Horatii
Leonardo da Vinci	Alexandros of An...		Eugène Delacroix	Antonio Canova	Théodore Géricault	Jacques-Louis D...	Paolo Veronese		Leonardo da Vinci	Leonardo da Vinci	Jacques-Louis D...

Feedback

Paris City Vision
https://www.pariscityvision.com › ... › Louvre museum

Louvre artwork : top masterpieces and paintings

How can we not mention the **Mona Lisa**? The portrait assumed to be of the wife of Francesco del Giocondo is considered to be the most famous painting in the world ...

People also ask

What is the most famous artwork in Louvre?

What are the three masterpieces of the Louvre?

What are the big 3 at the Louvre?

Where is the real Mona Lisa painting?

Feedback

Le Louvre
https://www.louvre.fr › explore › visitor-trails › the-lou...

The Louvre's Masterpieces - What exactly is a ...

The palace is home to some of the **world's most** iconic pieces – **paintings**, sculptures, architectural elements and **art** objects by **famous** or anonymous artists.

See photos · See outside

Louvre Museum

Website · Directions · Save

4.7 ★★★★½ 325,915 Google reviews

Museum in Paris, France

SEE TICKETS

Sponsored

Withlocals

Withlocals Your Way! - Paris City Tour
$87 · 5.0 ★ (2.2K)

The Louvre, or the Louvre Museum, is a national art museum in Paris, France, and one of the most famous museums in the world. Wikipedia

Departments: Librairie-Boutique du Musée du Louvre

11

which pieces from the salon de refuses are now in the louvre

All　Images　Videos　Shopping　News　Maps　Web　More　Tools

Sign in

Wikipedia
https://en.wikipedia.org › wiki › Salon_des_Refusés

Salon des Refusés

Today, by extension, salon des refusés refers to any exhibition of works rejected from a juried art show.

People also ask

Where is the Salon Carre in the Louvre?

What happened with the works entered in the Salon of the Refused?

Which painting was included in the first Salon des Rejectés Salon of the Rejected?

Does the Salon in Paris still exist?

Feedback

Artland Magazine
https://magazine.artland.com › Articles and Features

Contemporary Art History: The Salon Des Refusés

Discover the 1863 **Salon des** Refusés: first of a string of landmark contemporary art shows that have radically changed the course of Art History.

Le Louvre
https://www.louvre.fr › explore › visitor-trails › the-lou...

The Louvre's Masterpieces - What exactly is a ...

The palace is home to some of the world's most iconic **pieces** – **paintings**, **sculptures**, architectural elements and art objects by famous or anonymous artists.

The Tour Guy
https://thetourguy.com › ... › France › Paris › Louvre
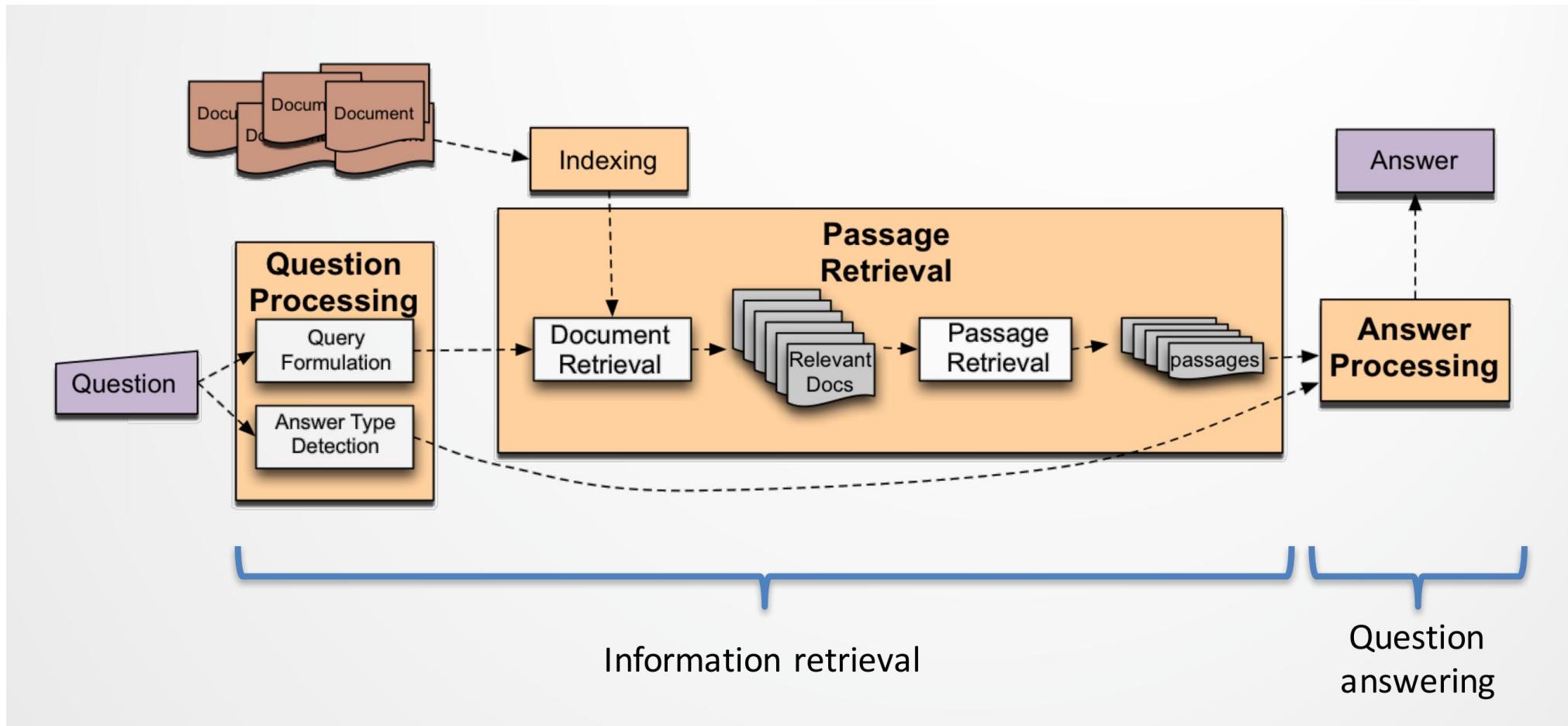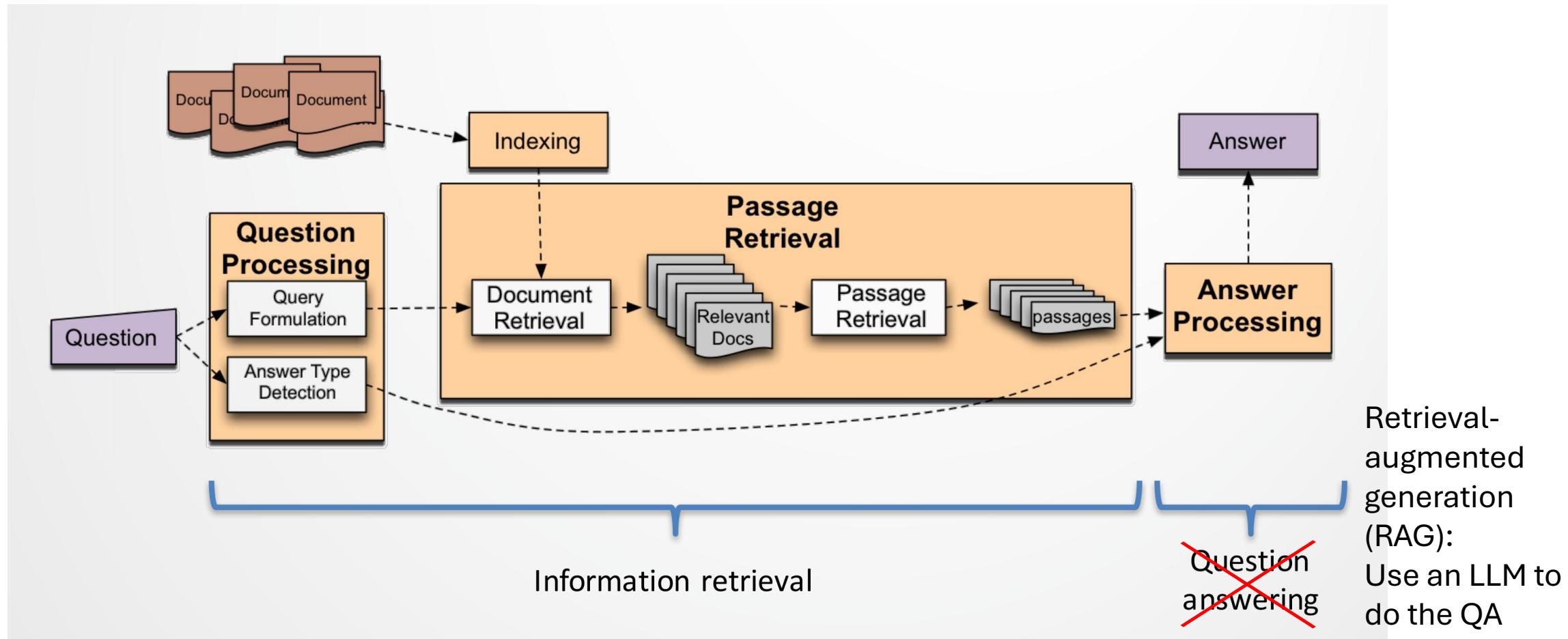
The Louvre Museum's 17 Most Important Works of Art, Paris

Oct 28, 2024 — The **Louvre** is massive. To make things easy, we've listed 17 famous **paintings** to see in the **Louvre** and explained why they're so important.

12

IR-based QA

Information retrieval

Question answering

IR-based QA with LLM (RAG)

Retrieval-augmented generation (RAG):
Use an LLM to do the QA

Sample TREC questions

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Qintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?
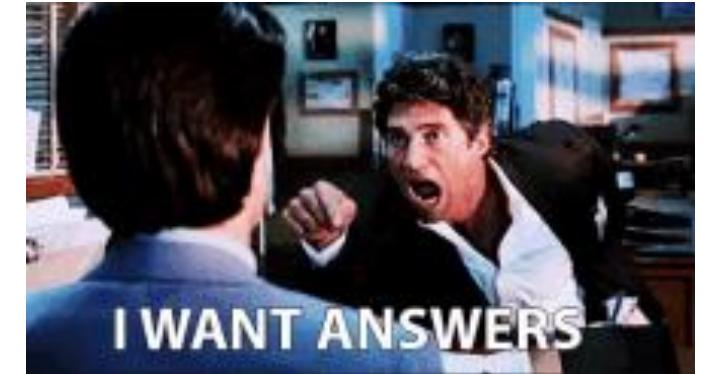
Query types

- Different kinds of questions can be asked.
    - Factoid questions, e.g.,
        - *How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?*
    - Narrative (open-ended) questions, e.g.
        - *Can you tell me about contemporary interest in the Greek philosophy of stoicism?*
    - Complex/hybrid questions, e.g.,
        - *Who was involved in the Schengen agreement to eliminate border controls in Western Europe and what did they hope to accomplish?*
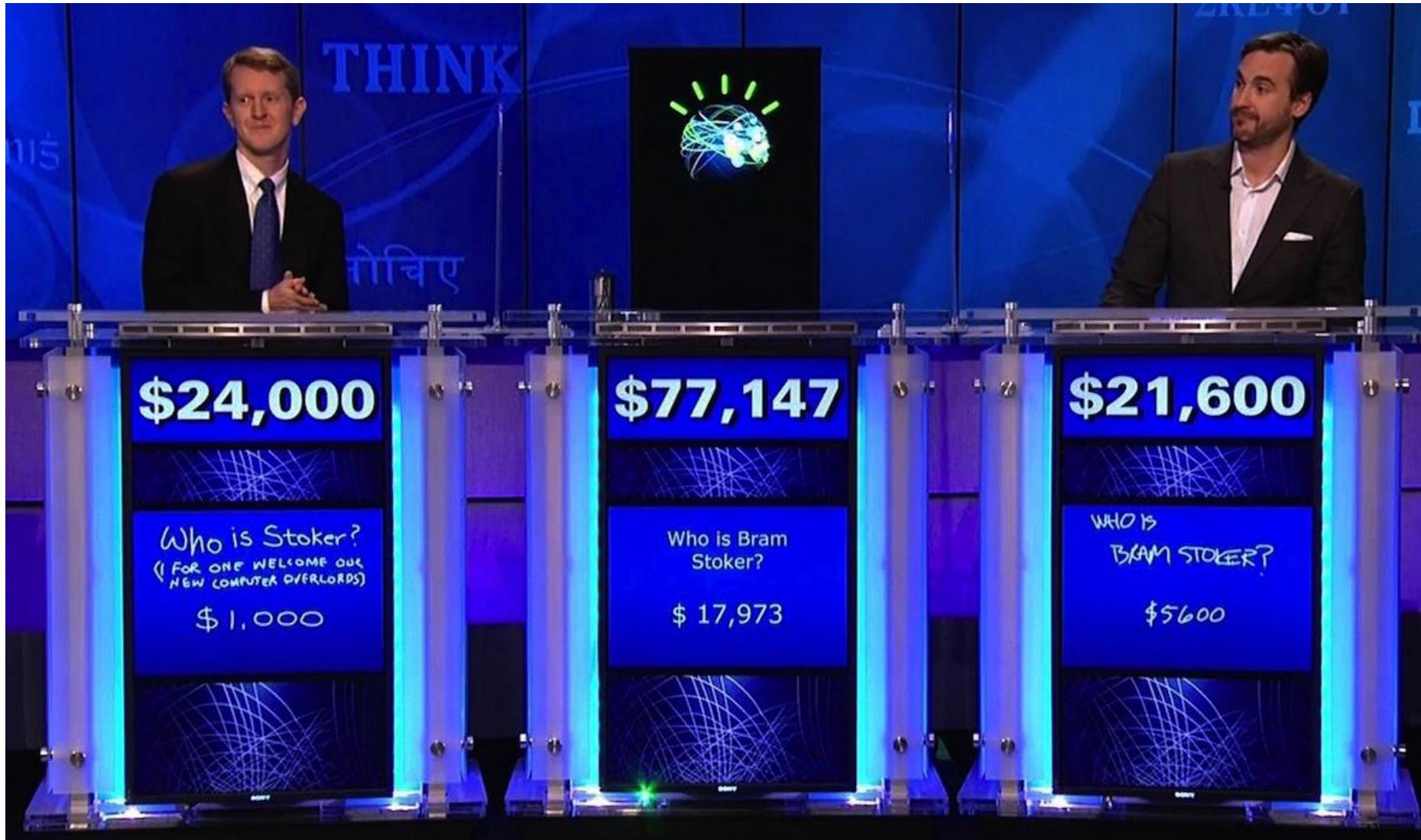
People **want** to ask questions…

I WANT ANSWERS

- **Examples from AltaVista query log (late 1990s)**
  - who invented surf music?
  - how to make stink bombs
  - where are the snowdens of yesteryear?
  - which english translation of the bible is used in official catholic liturgies?
  - how to do clayart
  - how to copy psx
  - how tall is the sears tower?
- **Examples from Excite query log (12/1999)**
  - how can i find someone in texas
  - where can i find information on puritan religion?
  - what are the 7 wonders of the world
  - how can i eliminate stress
  - What vacuum cleaner does Consumers Guide recommend
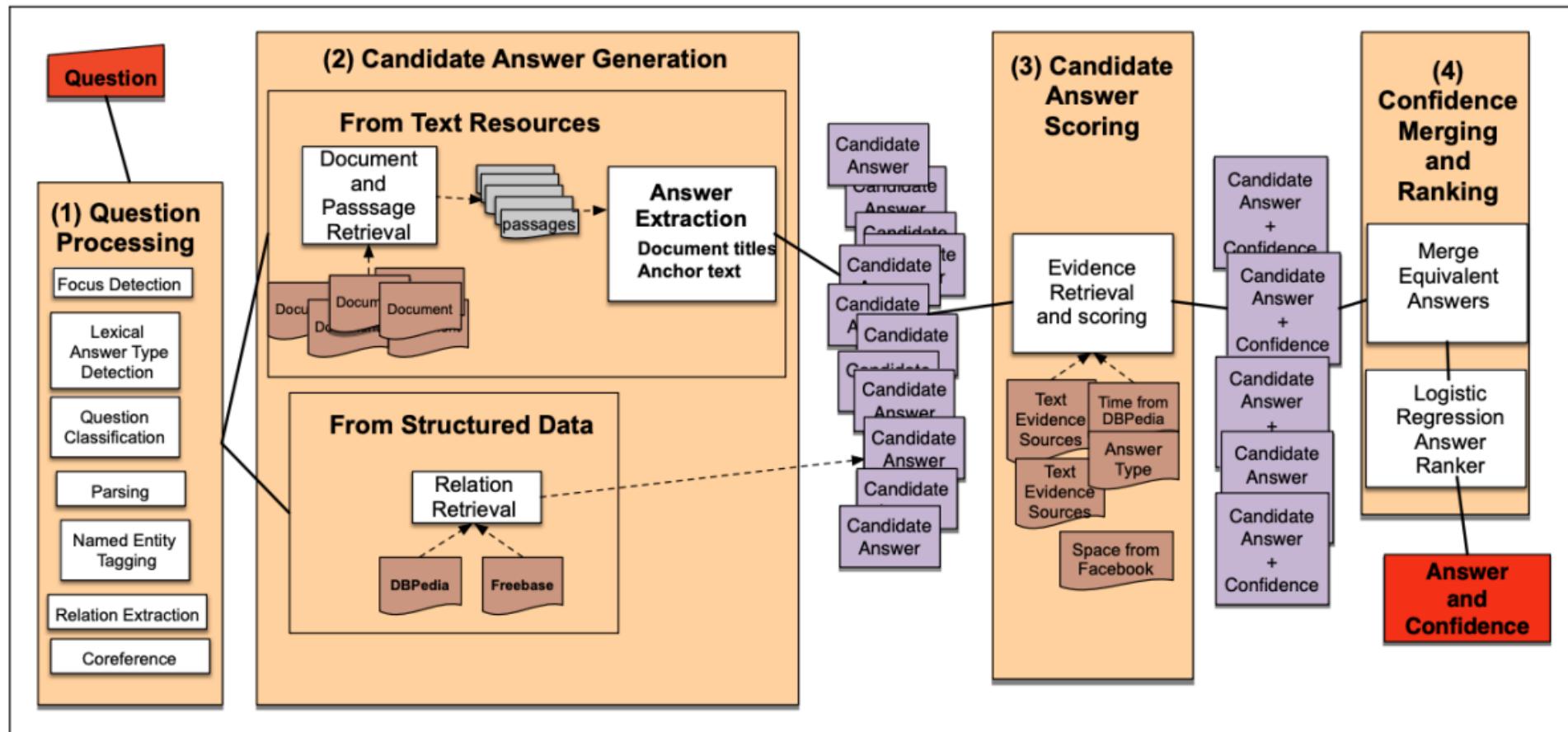
Around 10% of early query logs are QUESTIONS.

2011: IBM Watson beat Jeopardy champions
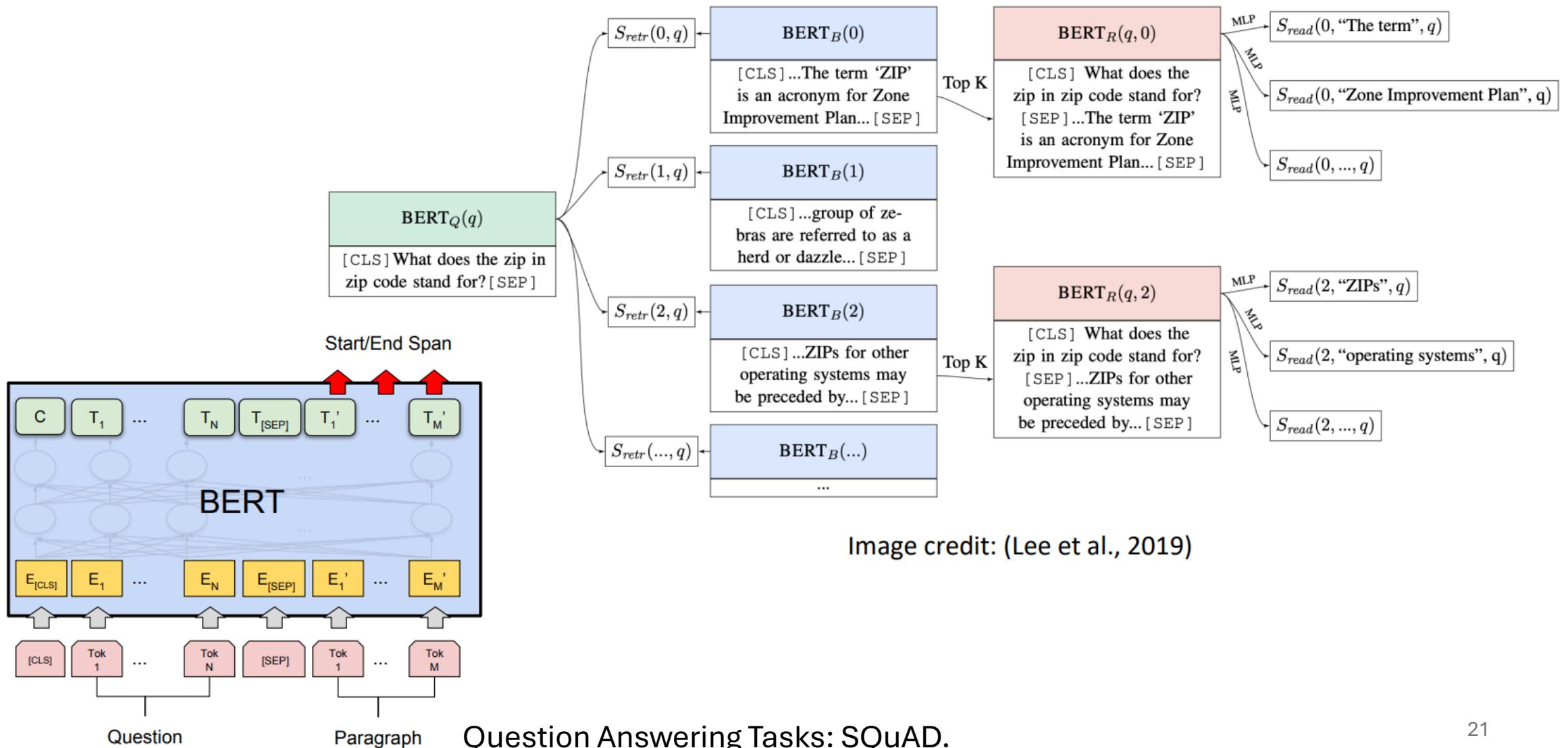
IBM Watson: Search

Not all problems are solved by these

- Where do lobsters like to live?
  - on a Canadian airline
- Where are zebras most likely found?
  - near dumps
  - in the dictionary
- Why can't ostriches fly?
  - Because of American economic sanctions
- What's the population of Mexico?
  - Three
- What can trigger an allergic reaction?
  - ..something that can **trigger** an allergic reaction

Question answering in deep learning era

Start/End Span

BERT

Question   Paragraph

$S_{retr}(0, q)$ ← BERT$_B$(0)
[CLS]...The term 'ZIP' is an acronym for Zone Improvement Plan...[SEP]

$S_{retr}(1, q)$ ← BERT$_B$(1)
[CLS]...group of zebras are referred to as a herd or dazzle...[SEP]

$S_{retr}(2, q)$ ← BERT$_B$(2)
[CLS]...ZIPs for other operating systems may be preceded by...[SEP]

$S_{retr}(..., q)$ ← BERT$_B$(...)
...

BERT$_Q(q)$
[CLS]What does the zip in zip code stand for?[SEP]

Top K

BERT$_R(q, 0)$
[CLS] What does the zip in zip code stand for? [SEP]...The term 'ZIP' is an acronym for Zone Improvement Plan...[SEP]

MLP → $S_{read}(0, \text{"The term"}, q)$
MLP → $S_{read}(0, \text{"Zone Improvement Plan"}, q)$
MLP → $S_{read}(0, ..., q)$

Top K

BERT$_R(q, 2)$
[CLS] What does the zip in zip code stand for? [SEP]...ZIPs for other operating systems may be preceded by...[SEP]

MLP → $S_{read}(2, \text{"ZIPs"}, q)$
MLP → $S_{read}(2, \text{"operating systems"}, q)$
MLP → $S_{read}(2, ..., q)$

Image credit: (Lee et al., 2019)

Question Answering Tasks: SQuAD.

21

SQuAD: Stanford question answering dataset

- 100k annotated (passage, question, answer) triples
  - Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!

- Passages are selected from English Wikipedia, usually 100~150 words.

- Questions are crowd-sourced.

- Each answer is a short segment of text (or span) in the passage.
  - This is a limitation— not all the questions can be answered in this way!

- SQuAD was for years the most popular reading comprehension dataset; it is "almost solved" today (though the underlying task is not,) and the state-of-the-art exceeds the estimated human performance.

- SQuAD 2.0: some questions can't be answered.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".
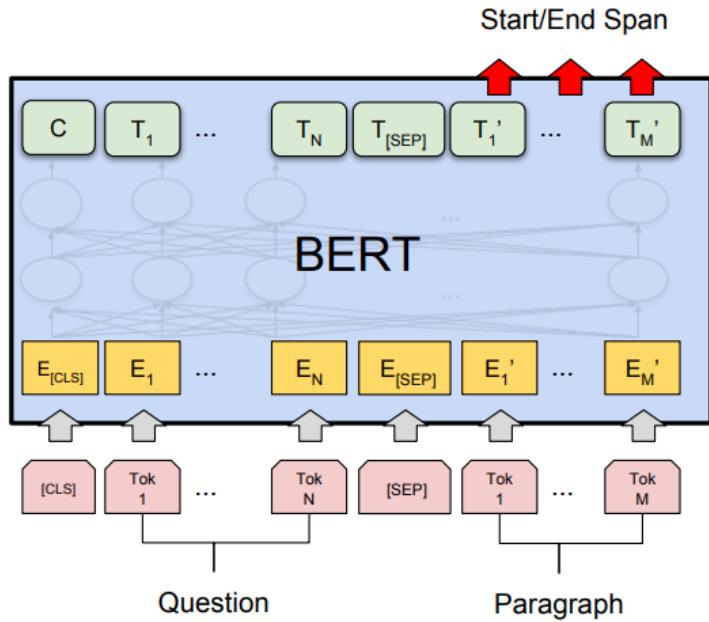
What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel
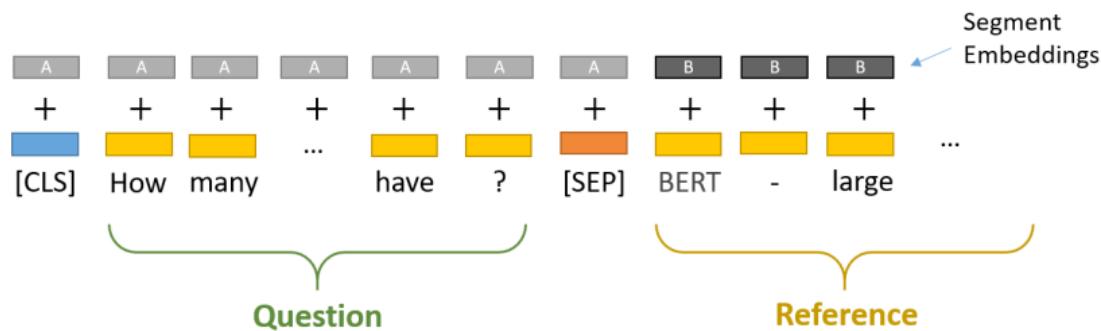
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

https://rajpurkar.github.io/SQuAD-explorer/

BERT for Reading Comprehension

Start/End Span

BERT

Question   Paragraph

- This simplified version of QA aka **Reading Comprehension**.
  - (Passage, Question) ⇒ Answer

Segment Embeddings

[CLS] How many ... have ? [SEP] BERT - large ...
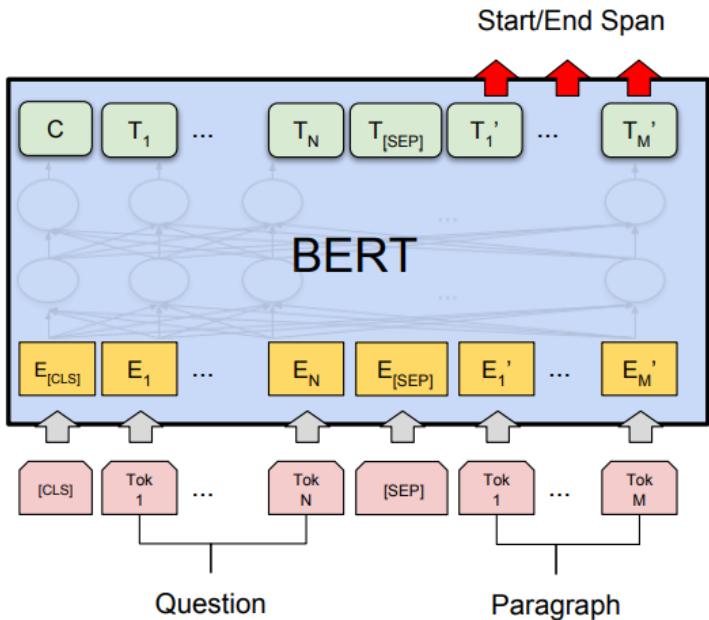
Question   Reference

Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

$$p_{\text{start}}(i) = \text{softmax}(\mathbf{W}_s \mathbf{h}_i)$$
$$p_{\text{end}}(i) = \text{softmax}(\mathbf{W}_e \mathbf{h}_i)$$

BERT for Reading Comprehension

Start/End Span

BERT

Question | Paragraph
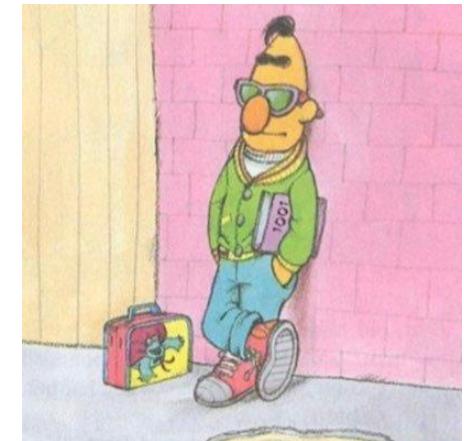
$$p_{\text{start}}(i) = \text{softmax}(\mathbf{W}_s \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}(\mathbf{W}_e \mathbf{h}_i)$$
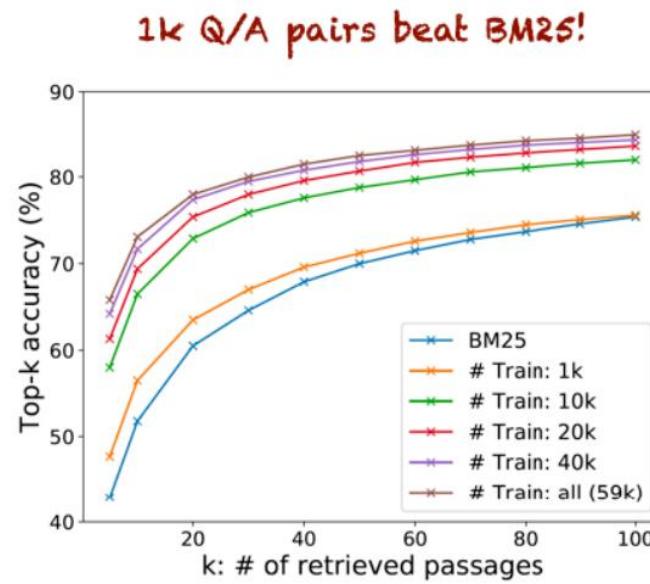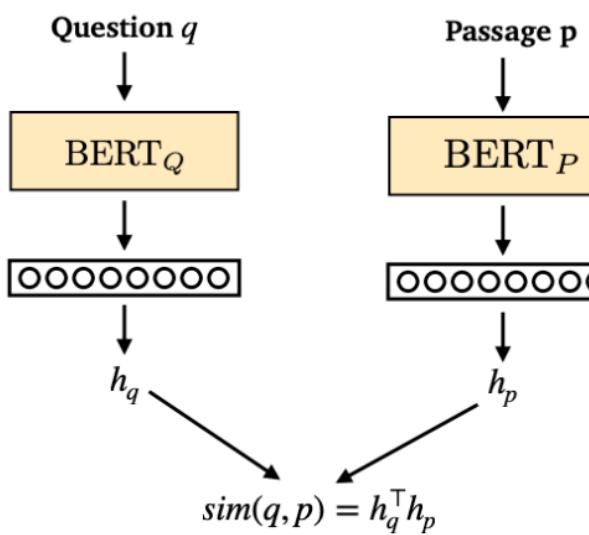
- This simplified version of QA aka **Reading Comprehension**.
  - (Passage, Question) ⇒ Answer

	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

BERT for IR

- Dense passage retrieval (DPR)
  - We can also just train the retriever using question-answer pairs!
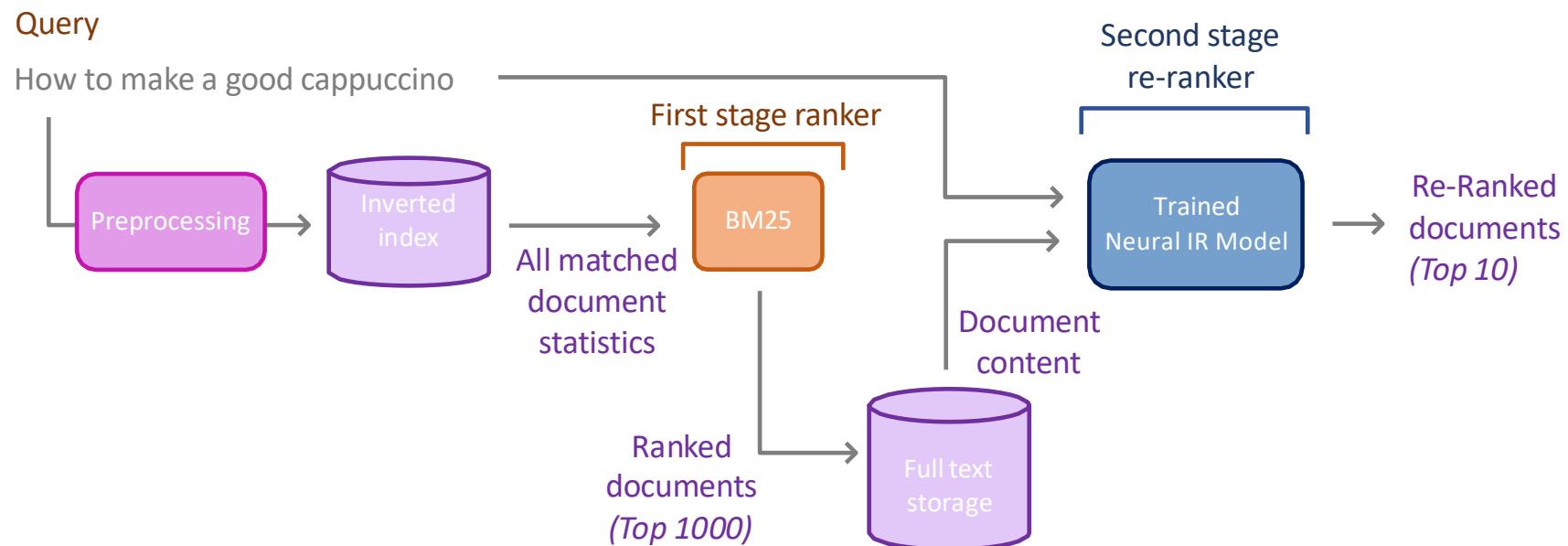
- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models.

Karpukhin et al., (2020). Dense Passage Retrieval for Open-Domain Question Answering.

Neural Methods for IR Beyond Re-ranking

- Re-Ranking depends on a candidate selection (bottleneck)
  - How to bring neural advances in this first-stage phase
- Today we look at dense retrieval as (inverted index) BM25 alternative
- Many other neural approaches to improve first-stage retrieval:
  - **Doc2query**: Document expansion with query text that would semantically match the document. Exists in both BERT and T5 variants. Then index the expanded documents with BM25
  - **DeepCT**: Assign term weights based on BERT output during indexing -> retrieval with inverted index & BM25
  - **COIL**: Fuses contextual vectors into an inverted index structure, for faster lookup of semantic matches
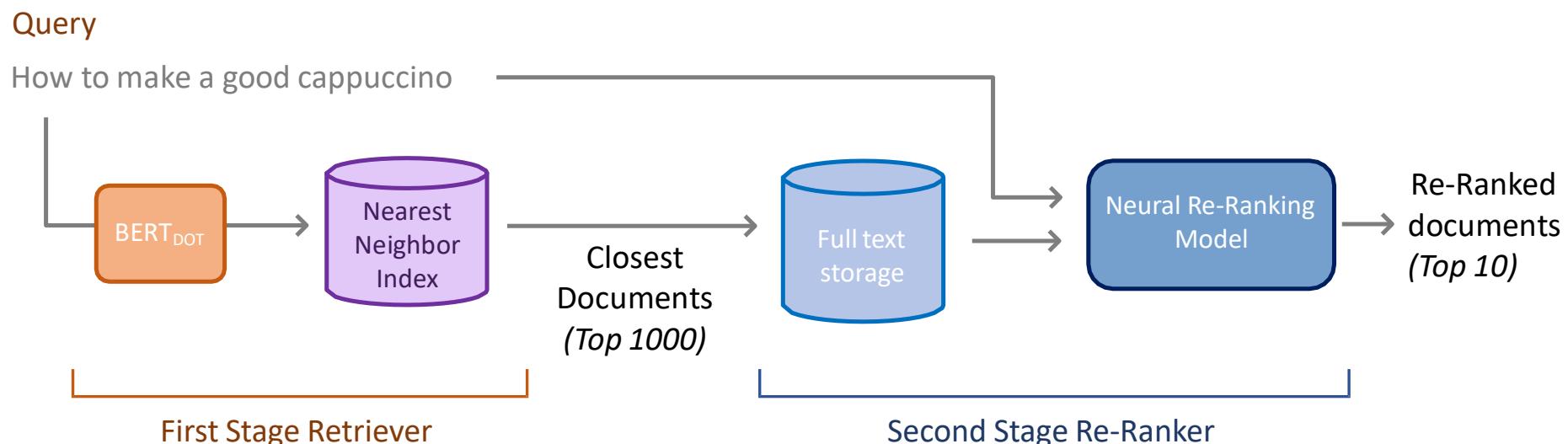
Neural Re-Ranking

- Re-rankers: They change the ranking of a pre-selected list of results
  - Same interface as classical ranking methods: score(q, d)
- Query workflow:

Dense Retrieval (with Re-Ranking)

- Dense retrieval replaces the traditional first stage
  - Using a neural encoder & nearest neighbor vector index
  - Can be used as part of a larger pipeline
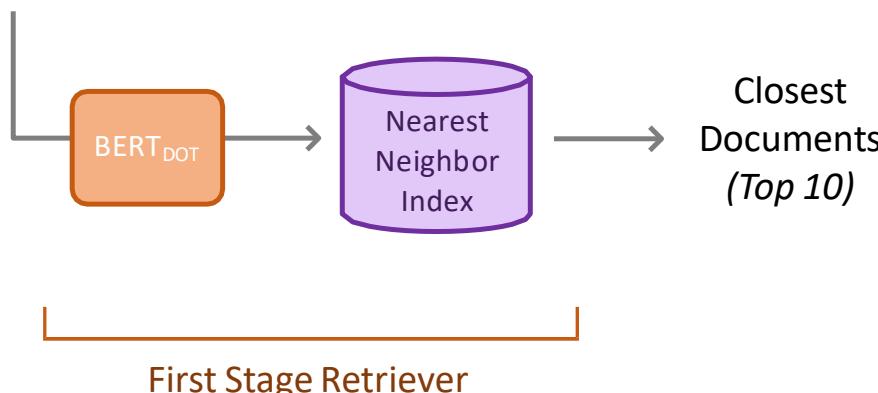
Standalone Dense Retrieval

- If dense retrieval is effective enough for our goals:
  - We can also use it as a standalone solution
  - Much faster + less complexity if we remove re-ranking stage

Query
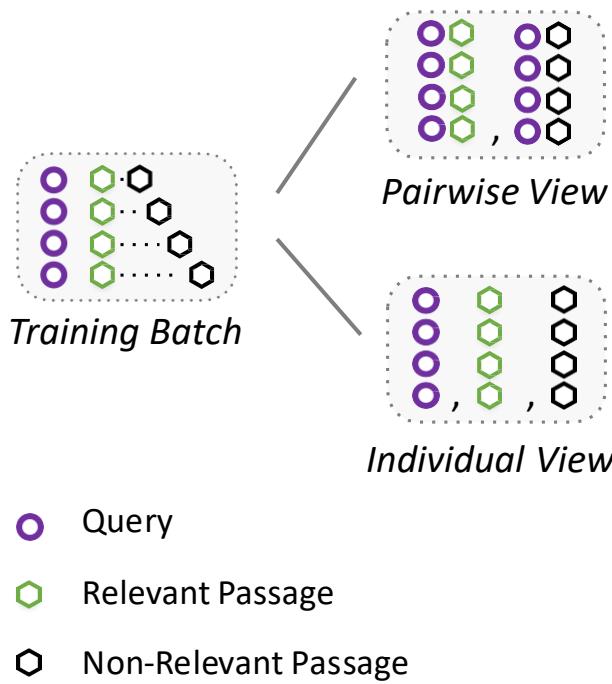
How to make a good cappuccino

BERT$_{DOT}$ → Nearest Neighbor Index → Closest Documents *(Top 10)*

First Stage Retriever

Training

- Neural IR models are typically trained with triples (pairwise +,-)
  - Triple: 1 query, 1 relevant, 1 non-relevant document
  - Generate embeddings for query, relevant doc, non-relevant doc
  - Loss function: Maximize margin between rel/non-rel document

- All model components are trained end-to-end
  - Of course we could decide to freeze some parts for more efficient training

Creating Training Batches

Pairwise View

Training Batch

Individual View

- Query
- Relevant Passage
- Non-Relevant Passage

- We form a batch by sampling as many triples as is allowed by the GPU memory
  - Typical batch size: 16-128
  - We mix different queries together
  - Depending on the model we need to create query-passage pairs or run each of the three sequences individually through the model

- We run a backward pass & gradient update per batch

- Sequency inputs come as a single matrix, so we need to pad different length inputs

Sampling Non-Relevant Passages

- Most collections only come with judgements of relevant (or false-positive selections from other models) and not truly non-relevant judgements
  - It doesn't make sense to spend resources annotating random pairs

- We need to tell the model what is non-relevant
  - Simple procedure to sample non-relevant passages:
    - Run BM25 and get the top-1000 results per training query and randomly select a few of those results as non-relevant
    - The non-relevant selections provide some signal
(as there must be at least some lexical overlap)
    - But mostly non-relevant passages -> works pretty good in practice
    - A bit of noise is good (we don't know the degree of non-relevance, but that's ok)

Loss Function
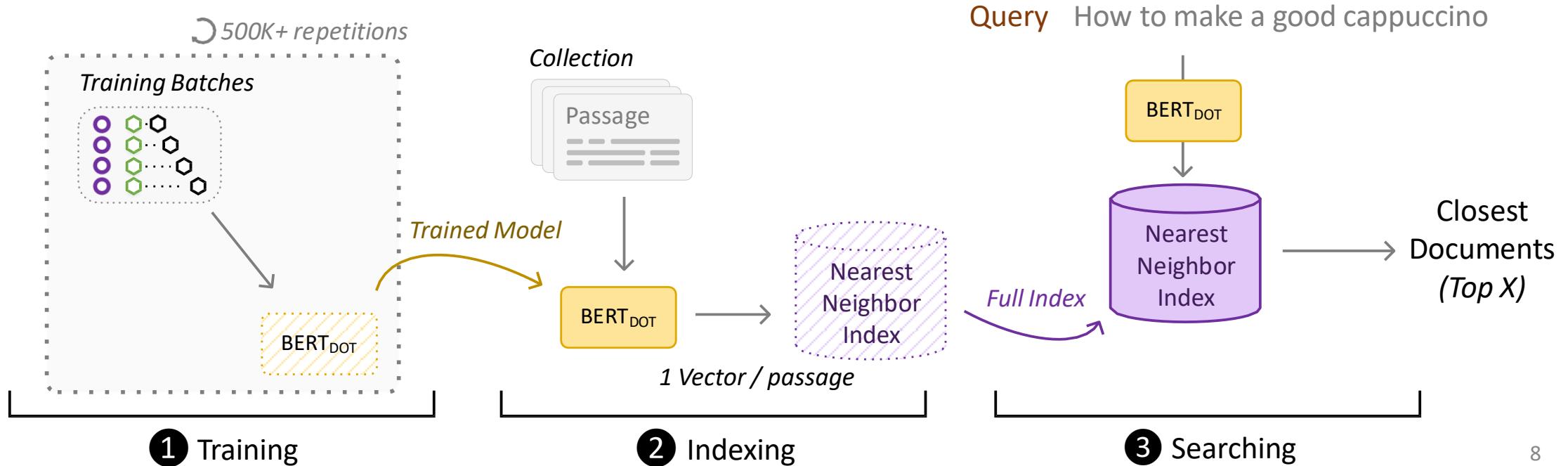
- Choice of different methods that aim to maximize the margin between rel & non-rel document
  - Plain Margin Loss:

    `loss = max(0,s_nonrel – s_rel + 1)`
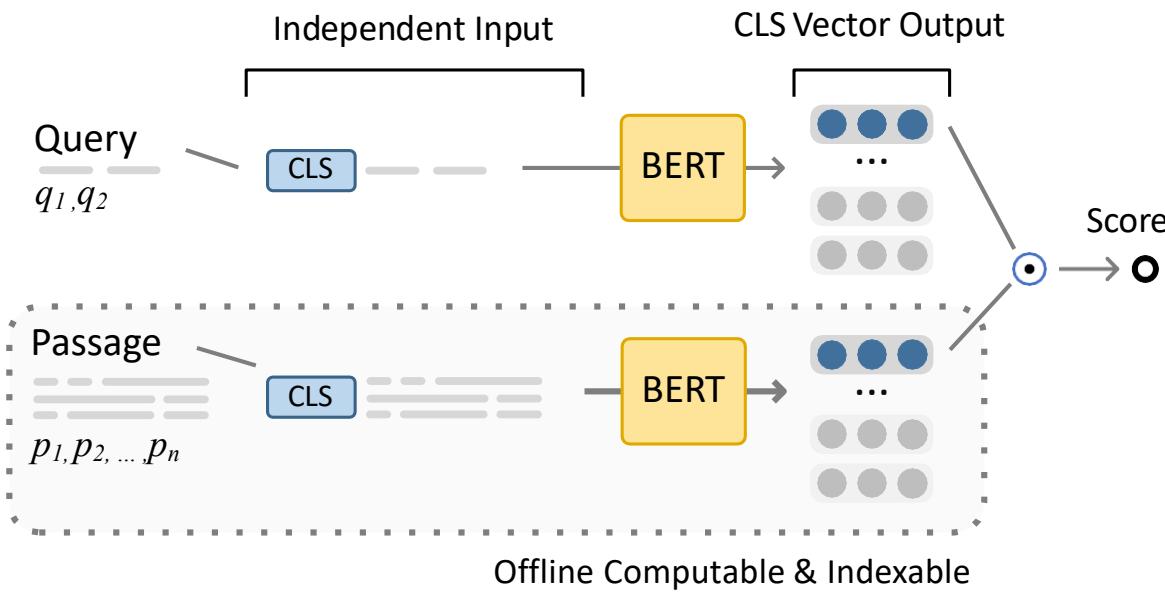
    - Native support in PyTorch: `torch.nn.MarginRankingLoss()`

  - RankNet:

    `loss = BinaryCrossEntropy(s_rel – s_nonrel)`

- Both losses assume binary relevance

Dense Retrieval Lifecycle

- 3 major phases in the dense retrieval lifecycle
  - Each comes with several complex choices and required techniques
  - Could skip ❶ if we use a pre-trained model

❶ Training   ❷ Indexing   ❸ Searching

BERT$_{DOT}$ Model

- Passages and queries are compressed into a single vector
  - Passages are completely independent -> moves most computation into the indexing phase
  - Only need query encoding at runtime

- Relevance is scored with a dot-product
  - Cosine-sim variants also exist
  - This allows easy use of an (approximate) nearest neighbor index

BERT$_{DOT}$

- Simple formula (as long as we abstract BERT):

$$\hat{q} = \text{BERT}([CLS]; q_{1..n})_{CLS}$$

$$\hat{p} = \text{BERT}([CLS]; p_{1..m})_{CLS}$$

Encoding

Independent computation

Matching

$$s = \hat{q} \cdot \hat{p}$$
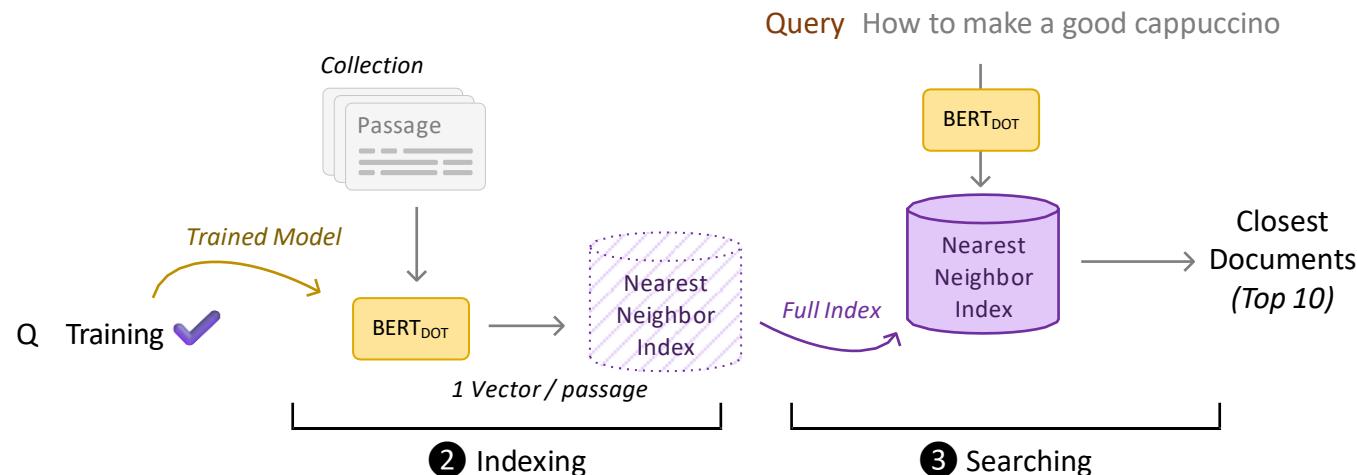
Can be done "outside" the model
(with a nearest neighbor library)

- Optional compression of $\hat{q}, \hat{p}$ with a single linear layer
(to reduce dimensionality)

$q_{1..n}$	Query tokens
$p_{1..m}$	Passage tokens
BERT	Pre-trained BERT model
$[CLS]$	Special tokens
x_{CLS}	Pool the CLS vector
s	Output score

Nearest Neighbor Search

- Once we have a trained DR model,
  we encode every passage in our collection
  - We save passages in an (approximate) nearest neighbor index
- During search we encode the query on the fly
  and search for nearest neighbor vectors in the passage index

NN Search: GPU Brute-Force

- Retrieving the top-1K from
  9 million vectors is fast
  - We need to do 9M dot-products (a
    very big matrix multiplication)
    with 768 dim. vectors
- GPUs are made for this
  - Vectors must fit in GPU memory
  - 70ms latency / query
  - Incredible scale when
    increasing the batch size
  - Using a CPU this takes ~1 sec. / q

Table 1: Latency analysis of Top-1000 retrieval using our BERT$_{DOT}$ retrieval setup for all MSMARCO passages using DistilBERT and Faiss (FlatIP) on a single Titan RTX GPU
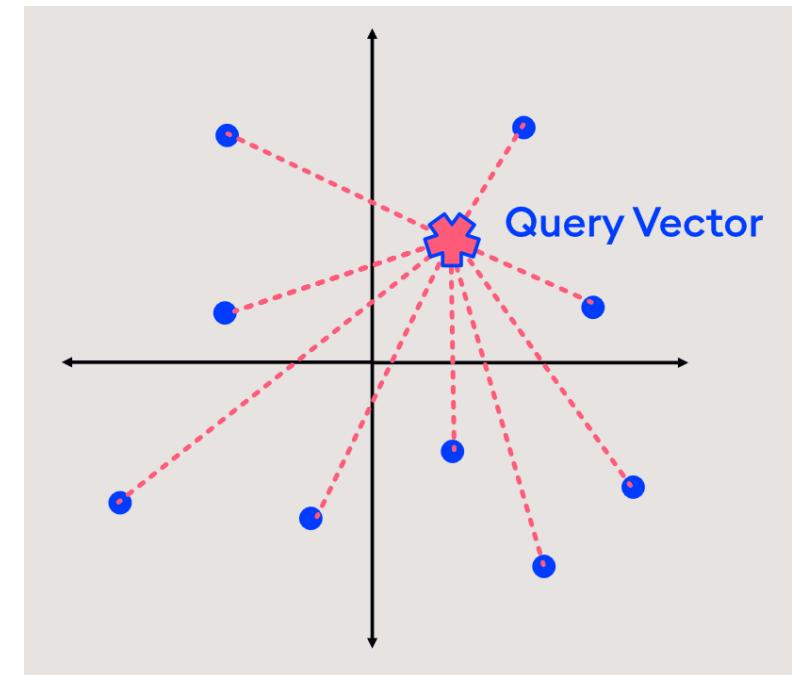
Batch Size	Q. Encoding		Faiss Retrieval		Total	
	Avg.	99th Per.	Avg.	99th Per.	Avg.	99th Per.
1	8 ms	11 ms	54 ms	55 ms	64 ms	68 ms
10	8 ms	9 ms	141 ms	144 ms	162 ms	176 ms
2,000	273 ms	329 ms	2,515 ms	2,524 ms	4,780 ms	4,877 ms

Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling; Hofstätter et al. SIGIR 2021
https://arxiv.org/abs/2104.06967

Indexing Techniques: Flat Index

Flat Index = Brute Force

- No additional processing, using raw vector embeddings

- Calculates distance for each pair, slow

- Exhaustive search, best accuracy

Indexing Techniques: Inverted File Index (IVF)
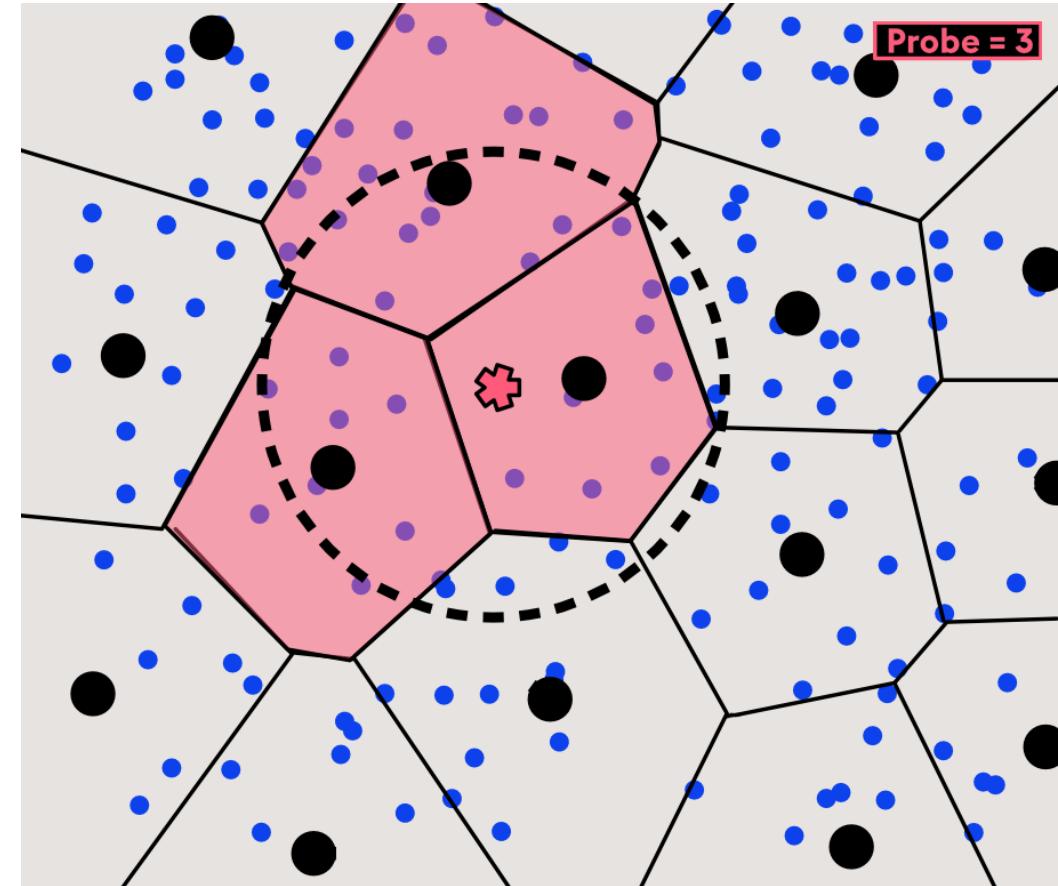
Partition the dataset into clusters

- Use clustering algorithm (e.g. k-means) to divide into k clusters

- Compute the centroids of each cluster

- For each cluster, store:

    - The centroid vector

    - An inverted index list of the vectors assigned to that cluster

Indexing Techniques: Inverted File Index (IVF)
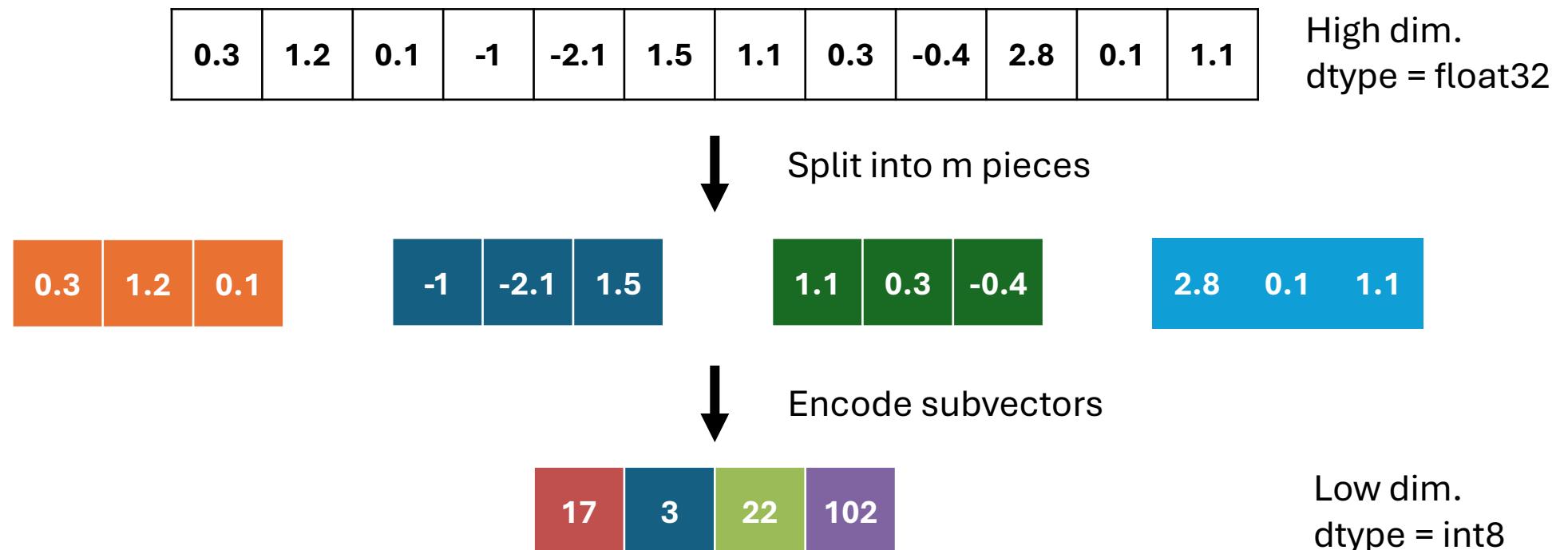
Query time:

- Compute similarity between query and centroids

- Select top n clusters

- Compute similarity to all vectors of these clusters

- Large reduction of search space

- More overhead
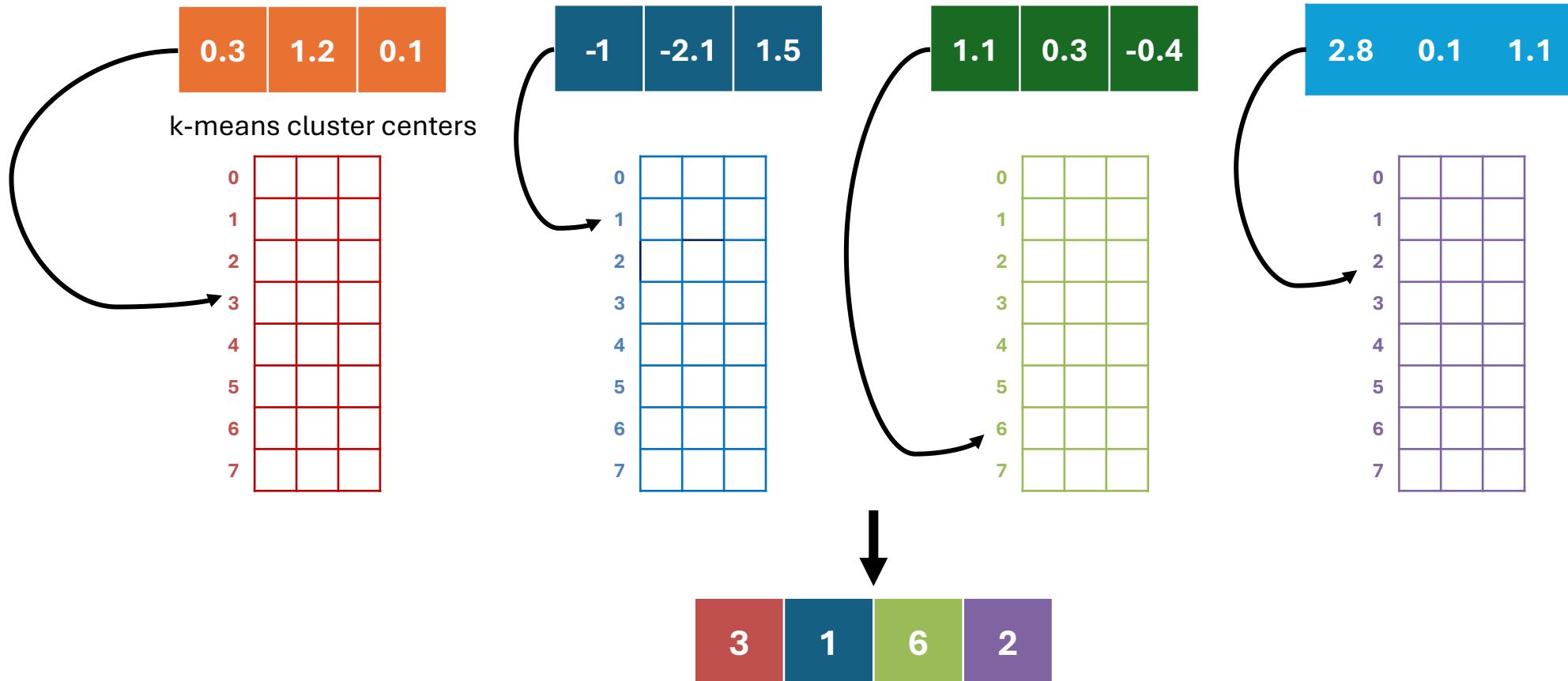
Indexing Techniques: Product Quantization

Main idea: Replace original vector of floats with lower dimensional vector of integers
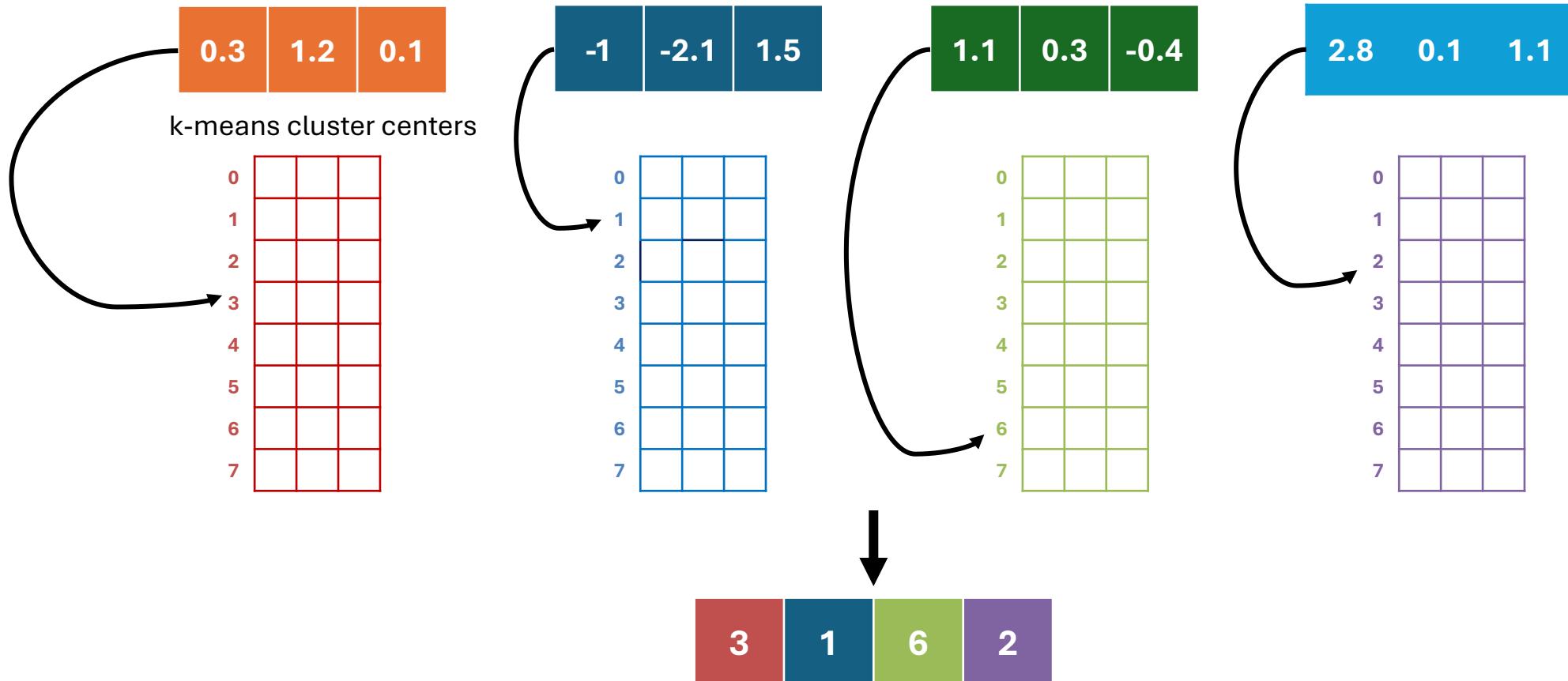
| 0.3 | 1.2 | 0.1 | -1 | -2.1 | 1.5 | 1.1 | 0.3 | -0.4 | 2.8 | 0.1 | 1.1 |

High dim.
dtype = float32

Split into m pieces

| 0.3 | 1.2 | 0.1 |

| -1 | -2.1 | 1.5 |

| 1.1 | 0.3 | -0.4 |

| 2.8 | 0.1 | 1.1 |

Encode subvectors

| 17 | 3 | 22 | 102 |

Low dim.
dtype = int8

Indexing Techniques: Product Quantization

Use k-means clustering on each sub-space

0.3	1.2	0.1

k-means cluster centers

-1	-2.1	1.5

1.1	0.3	-0.4

2.8	0.1	1.1

3	1	6	2

Indexing Techniques: Product Quantization

Use k-means clustering on each sub-space

| 0.3 | 1.2 | 0.1 |

k-means cluster centers

| -1 | -2.1 | 1.5 |

| 1.1 | 0.3 | -0.4 |

| 2.8 | 0.1 | 1.1 |

| 3 | 1 | 6 | 2 |

Indexing Techniques: Product Quantization

Reduces n * d (embedding space) matrix of floats to n * m integers

Additional preprocessing: Store distance from sub-vectors to centroids

Query time:

- Encode query vector in the same way
- Approximate distance from query to doc by sum of stored distance from doc sub-vector to query cluster centroid

Indexing Techniques: Product Quantization

Pros:

- Much faster

- Memory efficient
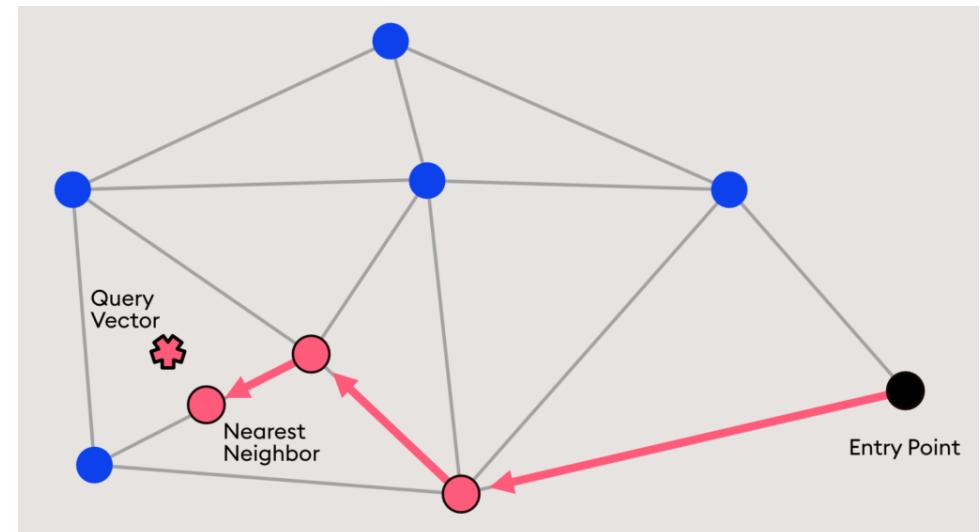
Cons:

- Results are approximate

- Quality depends on split and clustering parameters

Indexing Techniques: Graph Indices

e.g.: HNSW (Hierarchical Navigable Small Worlds)
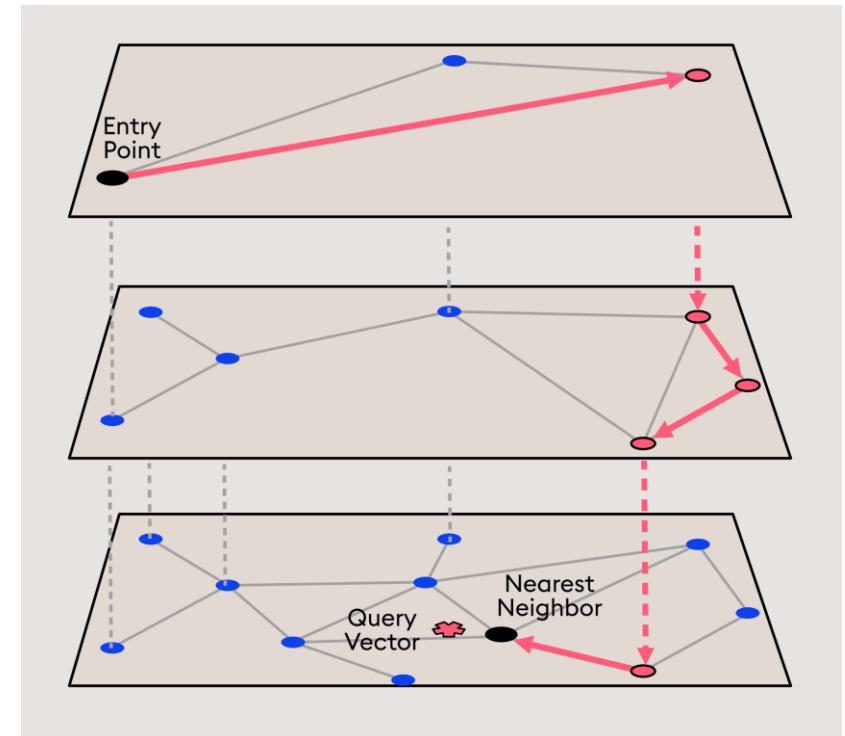
• Proximity graph, vectors are linked with similar "friends"

• Search starts at predefined "entry point",
visit "friends" until no nearer
vertex is found

Indexing Techniques: Graph Indices

Search space is split into hierarchical layers

- Top layer has longest distances

- When at a local minimum: drop one layer and keep searching

- Repeat until NN at lowest layer

- Needs additional pre-processing and memory, but scales much better to huge data sets

Approximate NN Search

- Brute-force search does not scale well beyond a couple of million vectors
- Fortunately, nearest neighbor search of vectors is a very common and broadly used technique in ML
  - many techniques and libraries to speed up search
- Popular library: FAISS
  - Offering many algorithms (brute-force, inverted lists on clusters, HNSW, …)
  - CPU and GPU supported
- Approximate search is another tradeoff between latency-effectiveness
  - We add a lot of complexity to the search system, but necessary for low-latency CPU serving

Production Support

- Dense retrieval is gaining more and more support in production systems
  - HuggingFace model hub gives us a common format to share models
  - Search engine must incorporate indexing & query encoding + provide nearest neighbor search
- Projects include Vespa.ai & Pyserini (integrates with Lucene)
  - Vespa provides deep integration of dense retrieval in common search features, such as filtering on properties
    - Important to filter during search, not after as to avoid empty result lists
  - Pyserini is a project focused on reproducing as many dense retrieval models as possible
    - Including easy hybrid search options between BM25 and DR
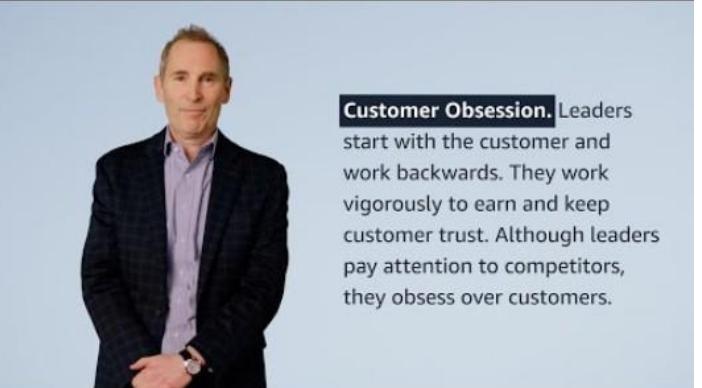
Other Uses for the BERT$_{DOT}$ Model

- Semantic comparisons of all sorts:
  - Sentence, passage, document similarity -> all compressed into 1 vector
  - Recommendation models

- S-BERT (Sentence transformers) library provides many models & scenarios
  - Based on the HuggingFace transformer library
  - Offers many scenarios and built models out of the box

- Adaptions based on dot-product similarity also allow for multi-modal comparisons
  - For example: Encoding images and text in the same vector-space

Next Lecture: RAG & Agent

NOW, ENJOY YOUR