344.175 VL: Natural Language Processing
Information Retrieval – Basics and Modern Approaches

Navid Rekab-saz

navid.rekabsaz@jku.at

Institute of Computational Perception

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

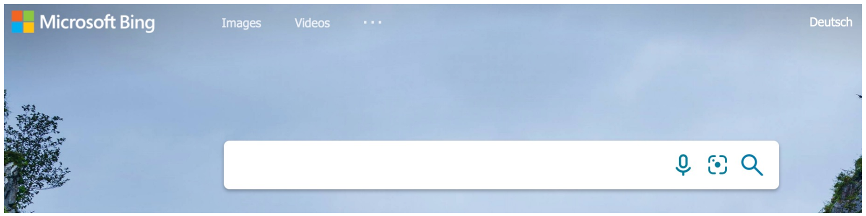
Agenda

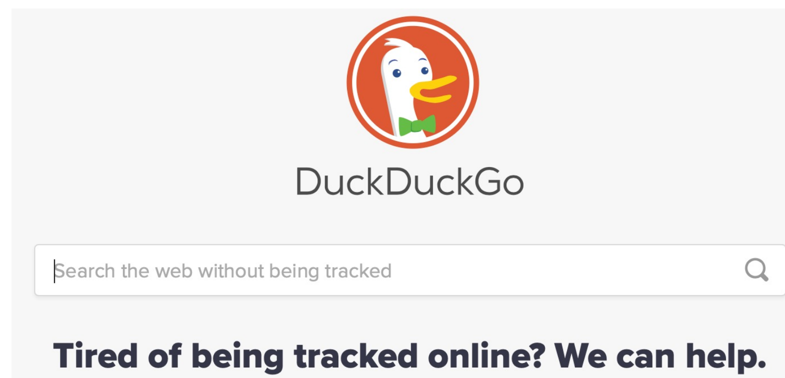
- Principles of Information Retrieval
- Evaluation of a ranked list
- IR with deep learning

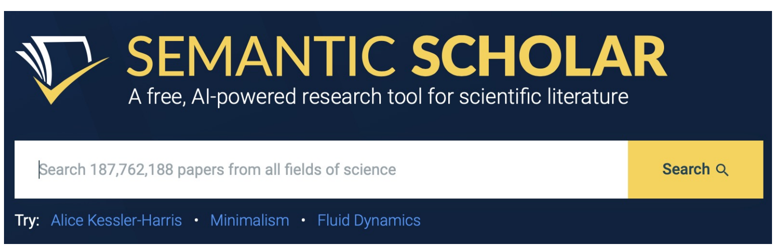
Agenda

- **Principles of Information Retrieval**
- Evaluation of a ranked list
- IR with deep learning

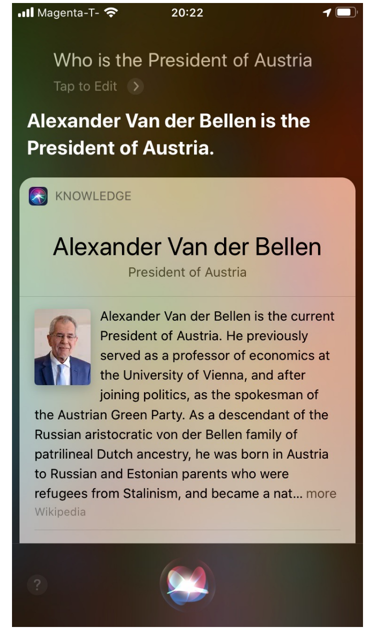
Information Retrieval everywhere!

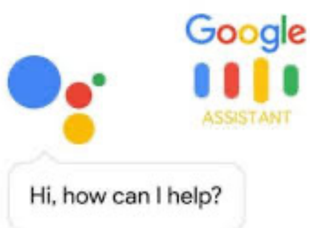






Information Retrieval everywhere!





IBM Watson and Jeopardy

Information Retrieval

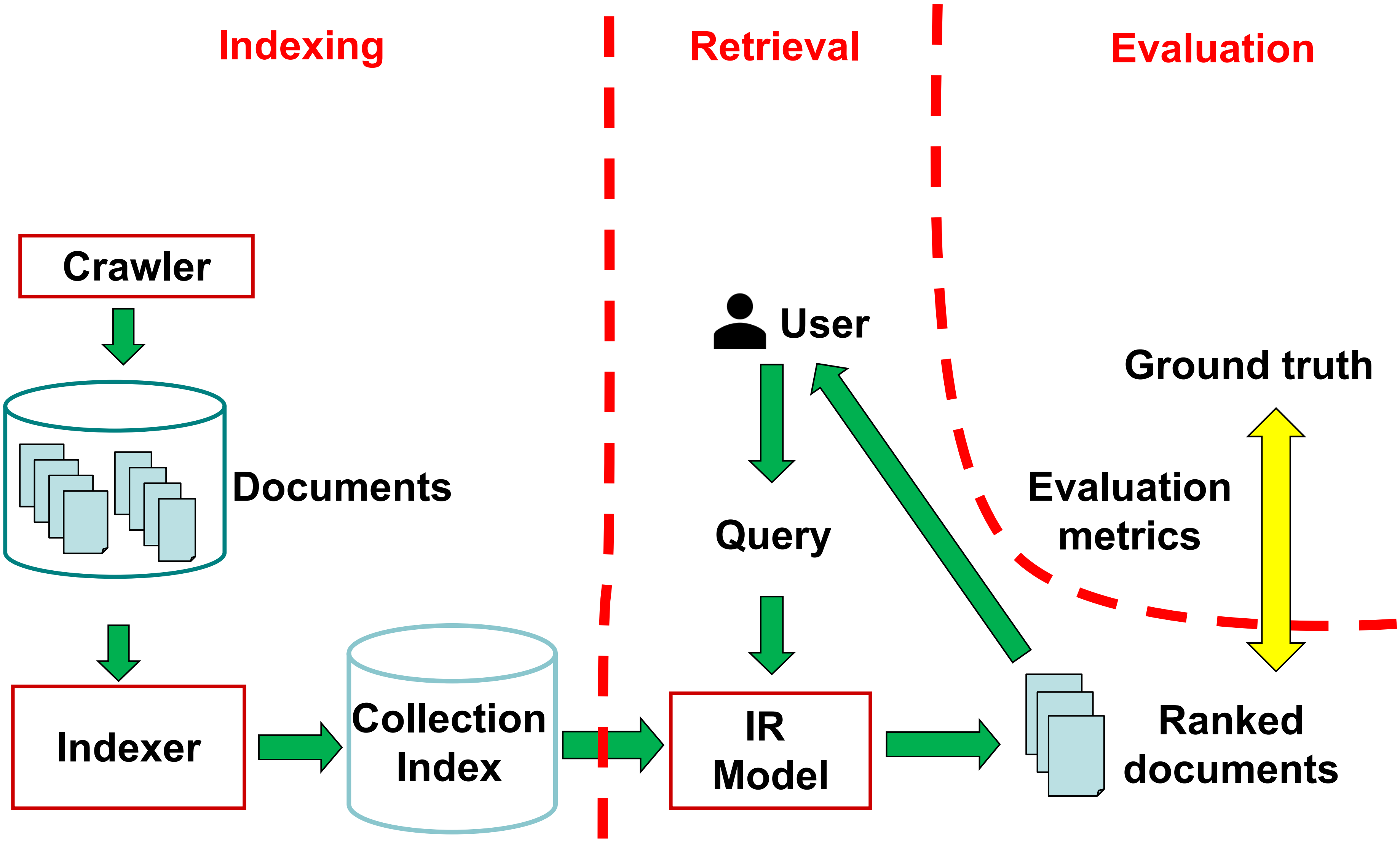
- Information Retrieval (IR) is finding material (usually in the form of documents) of an unstructured nature that satisfies an information need from within large collections

- When talking about IR, we frequently think of web search

- The goal of IR is however to retrieve relevant contents to the user's information need

- IR covers a wide set of tasks such as …
  - Ranking, question/answering, information summarization
  - But also … user behavior/experience study, personalization, etc.

Simplified architecture of an IR system



Terminology

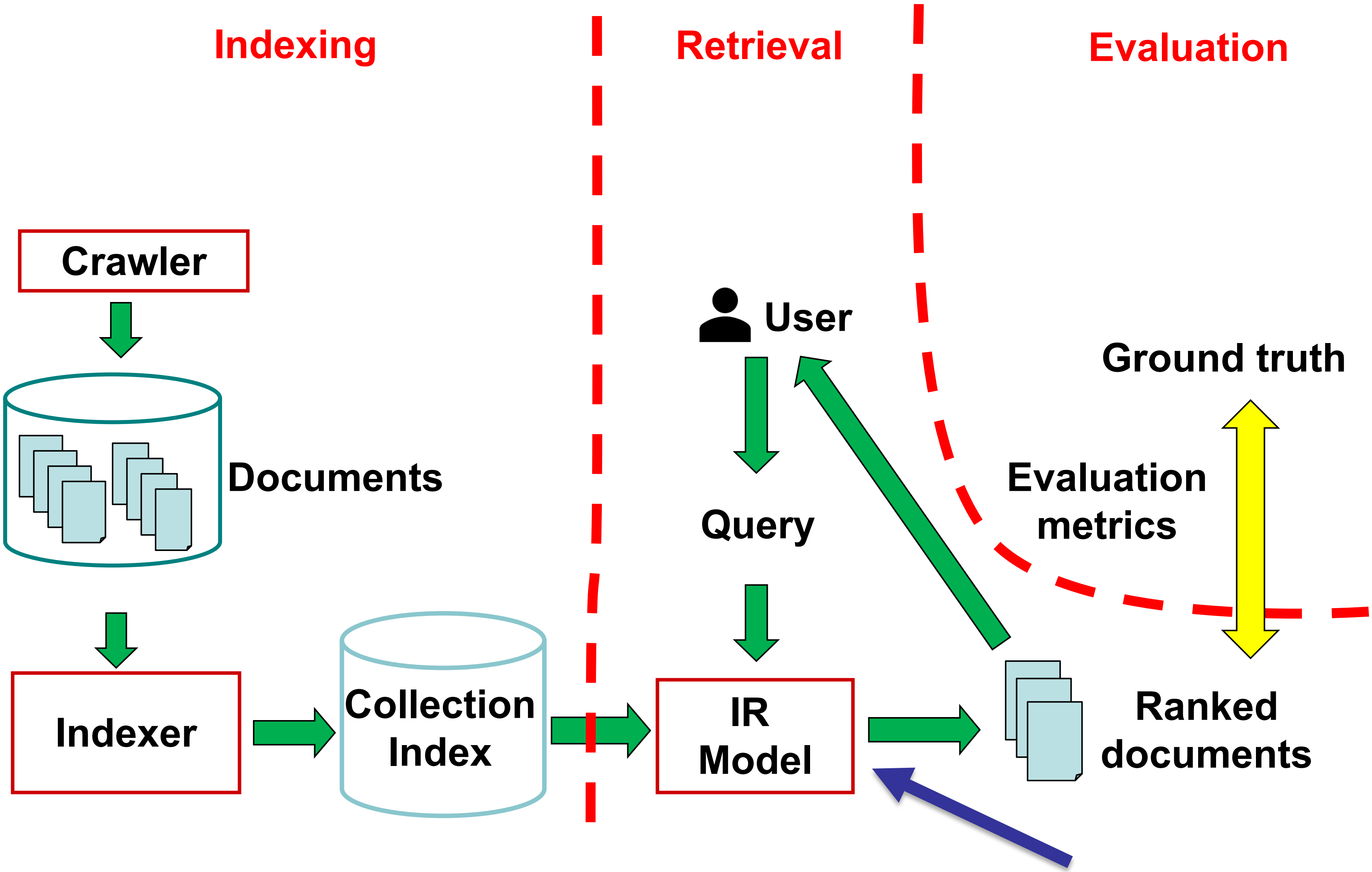
- Information need
  - E.g. *My swimming pool bottom is becoming black and needs to be cleaned*
- Query
  - A designed representation of users’ information need
  - E.g. *pool cleaner*
- Document
  - A unit of data in text, image, video, audio, etc.
- Relevance
  - Whether a document satisfies user’s information need
  - Relevance has multiple aspects: topical, semantic, temporal, spatial, etc.

Ad-hoc IR (all we discuss in this lecture)

- Studying the methods to estimate relevance, solely based on the contents (texts) of queries and documents
  - In ad-hoc IR, *meta-knowledge* such as temporal, spatial, user-related information are normally taken out
  - The focus of ad-hoc IR is on methods to exploit contents

- Ad-hoc IR is a part of the ranking mechanism of search engines, but there are several other aspects…
  - Diversity of information
  - Personalization
  - Information need understanding
  - Search engine log files analysis
  - …

Simplified architecture of an IR system



Relevance scoring & IR models

query: wisdom of mountains

(q)

$d20$

$d1402$

$d5$

$d100$

- Collection contains M documents. Each document d and each query q consists of a set of terms

- An IR model calculates/predicts a relevance score between the query and document:

$$s(q, d)$$

- Documents are ranked according to their predicted relevance scores to the query, from highest to lowest

Exact-matching IR models – TF-IDF

- Classical or exact-matching IR models – in their basic forms – assign importance weights to each query term that appears in a document

- <u>Recap:</u> TF-IDF was introduced as a term weighting method. It can also be an IR model to calculate relevance score:

$$s(q,d) = \sum_{q_i \in q} \text{tf-idf}_{q_i,d} = \sum_{q_i \in q} \underbrace{\boxed{\log\left(1 + \text{tc}_{q_i,d}\right)}}_{\textbf{Term matching score}} \times \underbrace{\boxed{\log\left(\frac{|\mathbb{D}|}{\text{df}_{q_i} + 1}\right)}}_{\textbf{Term Salience}}$$

$\text{tc}_{q_i,d}$ number of times query term q_i appears in document d

df_{q_i} number of documents in the collection, in which query term q_i appears

Exact-matching IR models – PL

- Pivoted Length Normalization model

Term matching score

$$s(q,d) = \sum_{q_i \in q} \frac{\log(1 + \mathrm{tc}_{q_i,d})}{1 - b + b\,\frac{|d|}{avgdl}} \times \mathrm{idf}\,(q_i)$$

Term Salience

Document length normalization

$\mathrm{tc}_{q_i,d}$ number of times query term q_i appears in document d

$avgdl$ average length of the documents in the collection

b a hyper parameter that controls document length normalization

Exact-matching IR models – BM25

- BM25 model *(slightly simplified)*:

Term matching score & normalization

$$s(q,d) = \sum_{q_i \in q} \frac{(k_1+1)\mathrm{tc}_{q_i,d}}{k_1\left(1-b+b\frac{|d|}{avgdl}\right)+\mathrm{tc}_{q_i,d}} \times \mathrm{idf}\,(q_i)$$

Term Salience

Length normalization

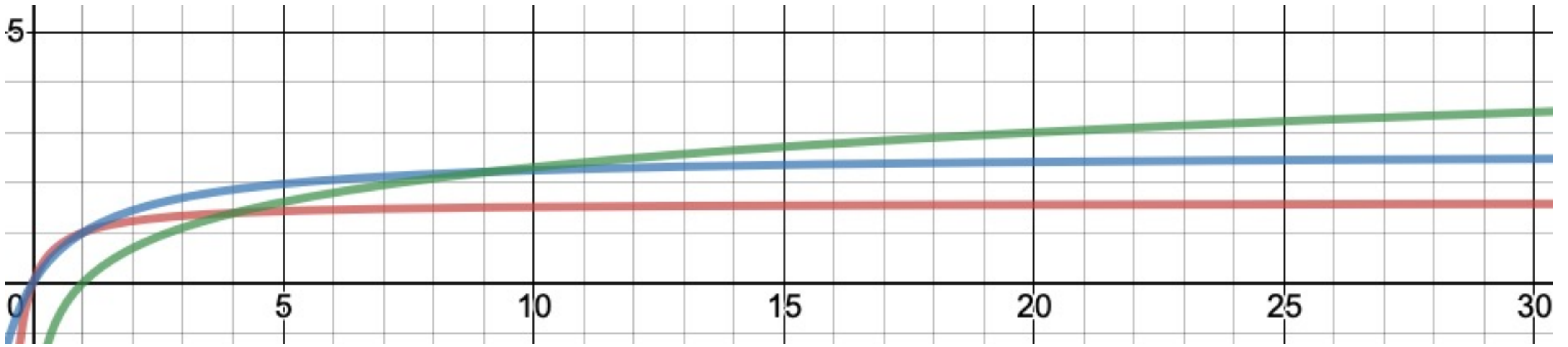
$\mathrm{tc}_{q_i,d}$ number of times query term q_i appears in document d

$avgdl$ average length of the documents in the collection

b a hyper parameter that controls length normalization

k_1 a hyper parameter that controls term frequency *saturation*

Exact-matching IR models – BM25

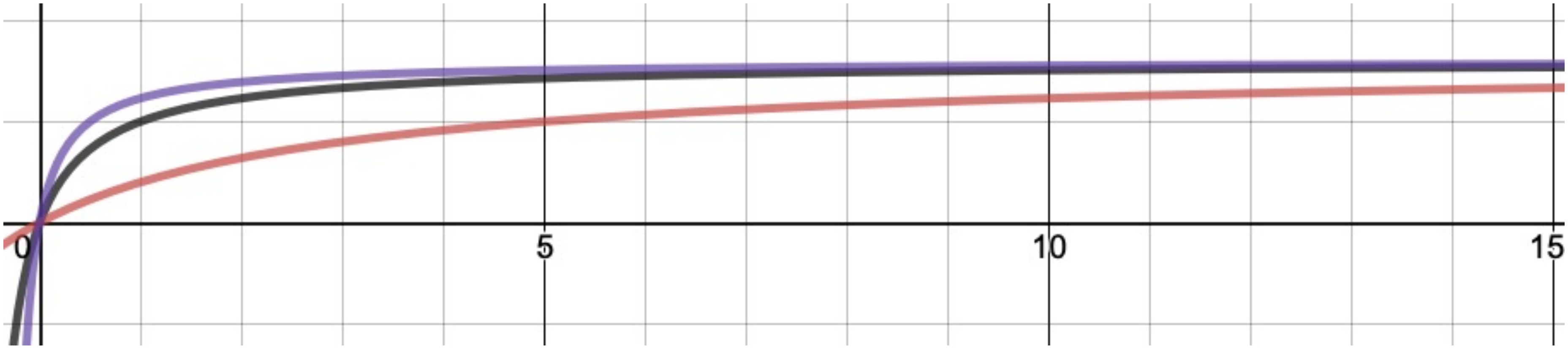


Green: $\log\ \mathrm{tc}_{q_i,d}$ → TF

Red: $\frac{(0.6+1)\mathrm{tc}_{q_i,d}}{0.6+\mathrm{tc}_{q_i,d}}$ → BM25 with $k_1 = 0.6$ and $b = 0$

Blue: $\frac{(1.6+1)\mathrm{tc}_{q_i,d}}{1.6+\mathrm{tc}_{q_i,d}}$ → BM25 with $k_1 = 1.6$ and $b = 0$

Exact-matching IR models – BM25



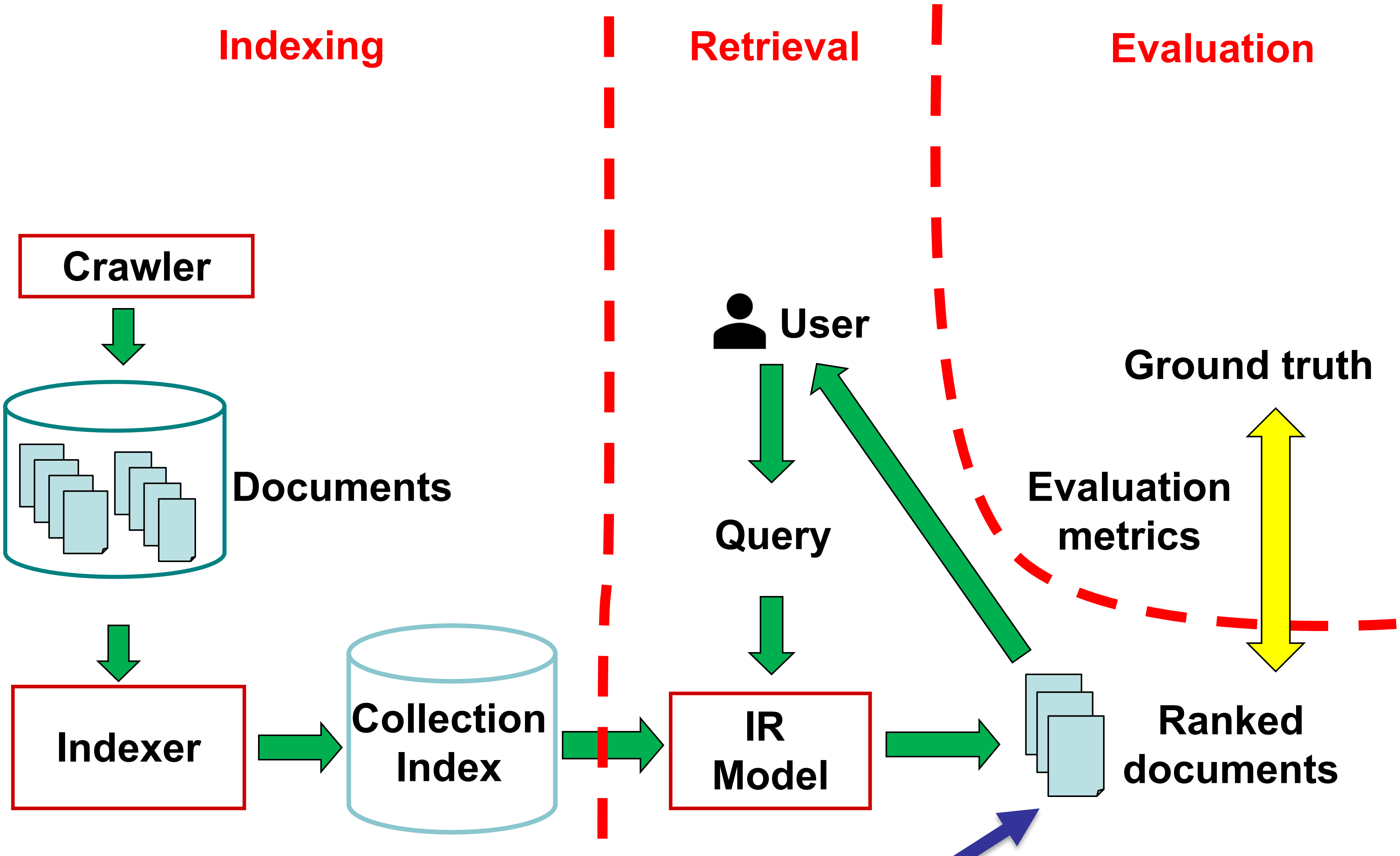
BM25 models with $k_1 = 0.6$ and $b = 1$

Purple: $\frac{(0.6+1)\mathrm{tc}_{q_i,d}}{0.6(1-1+1(\frac{1}{2}))+\mathrm{tc}_{q_i,d}}$ → Document length ½ of $avgdl$

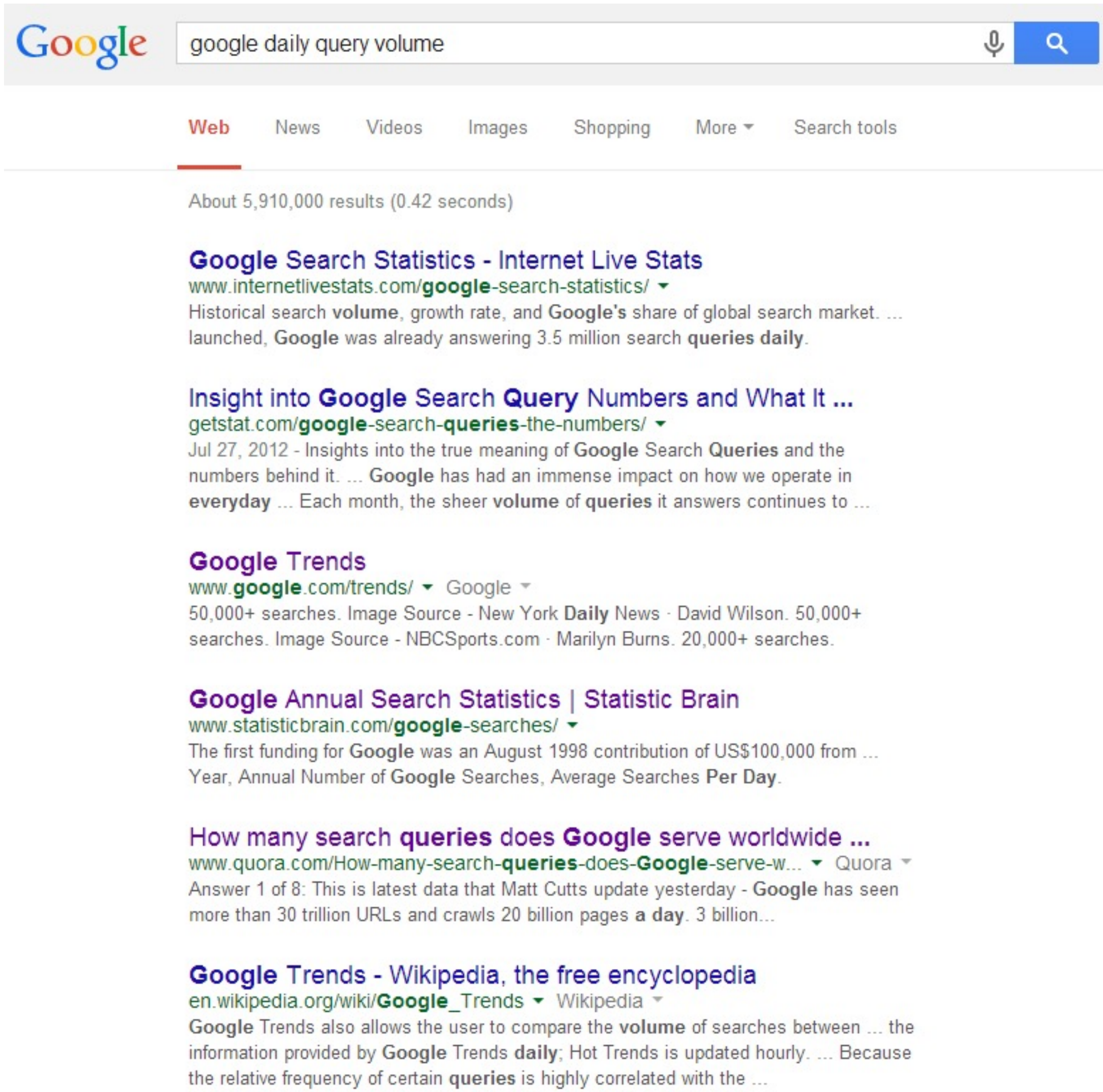
Black: $\frac{(0.6+1)\mathrm{tc}_{q_i,d}}{0.6(1-1+1(\frac{2}{2}))+\mathrm{tc}_{q_i,d}}$ → Document length the same as $avgdl$

Red: $\frac{(0.6+1)\mathrm{tc}_{q_i,d}}{0.6(1-1+1(\frac{10}{2}))+\mathrm{tc}_{q_i,d}}$ → Document length 5 times higher than $avgdl$

Simplified architecture of an IR system



Ranking results as we know!

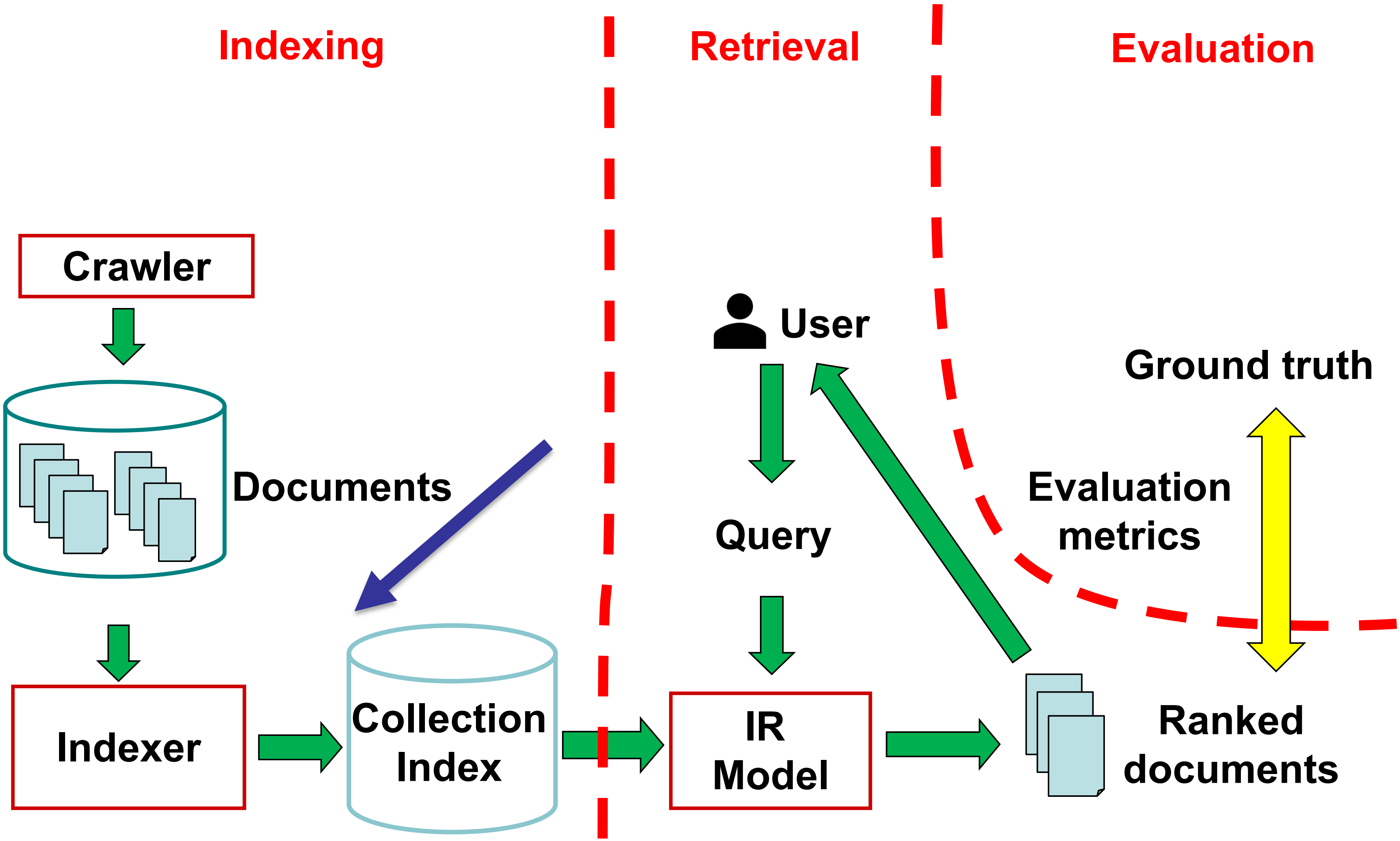


Sample ranking results – format in research!

- TREC run file: standard format to report the ranking results of top-1000 documents for some queries, retrieved by a model

qry_id	(iter)	doc_id	rank	score	run_id
2	Q0	1782337	1	21.656799	cool_model
2	Q0	1001873	2	21.086500	cool_model
			…		
2	Q0	6285819	999	3.43252	cool_model
2	Q0	6285819	1000	1.6435	cool_model
8	Q0	2022782	1	33.352300	cool_model
8	Q0	7496506	2	32.223400	cool_model
8	Q0	2022782	3	30.234030	cool_model
			…		
312	Q0	2022782	1	14.62234	cool_model
312	Q0	7496506	2	14.52234	cool_model
			…		

Simplified architecture of an IR system



Efficient retrieval with pre-computed Collection Index

query: wisdom of mountains
(q)

$d20$

$d1402$

$d5$

$d100$

How can we efficiently calculate relevance scores for documents? → Collection Index

☞ Since the IR models so far are based on exact matching, an we can focus on calculating relevance scores only for the documents that contain query terms → done by Inverted Index

Inverted index

- Inverted index is a data structure for efficient retrieval
  - Inverted index is created once at index time for all documents in the collection, and used for each query during query time

- Inverted index creates a posting list for each unique term in collection
  - A posting list of a term contains the list of the IDs of the documents, in which the term appears

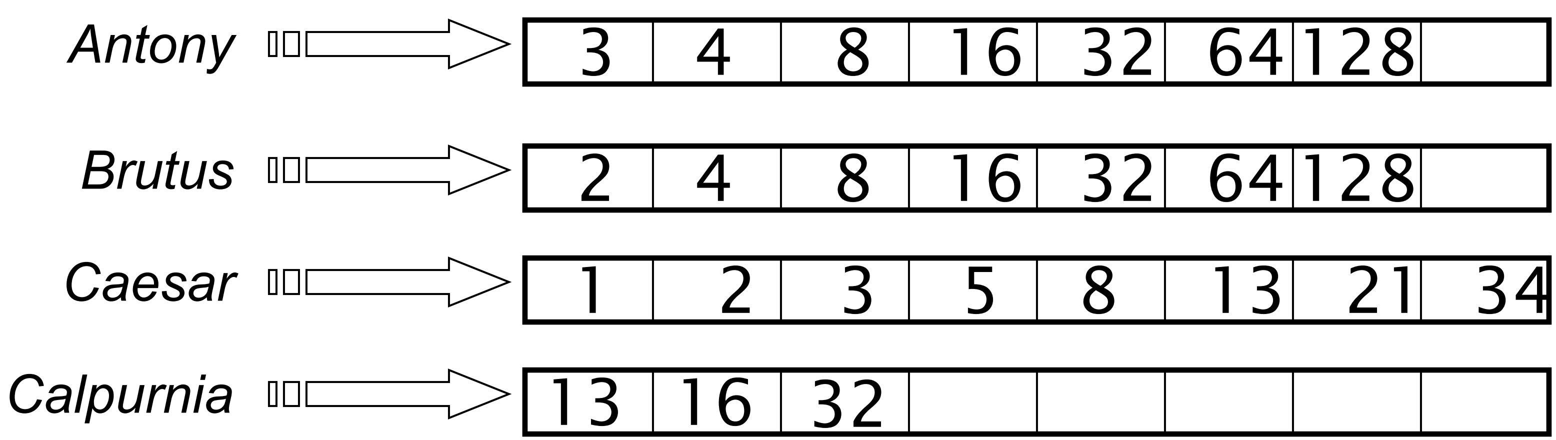
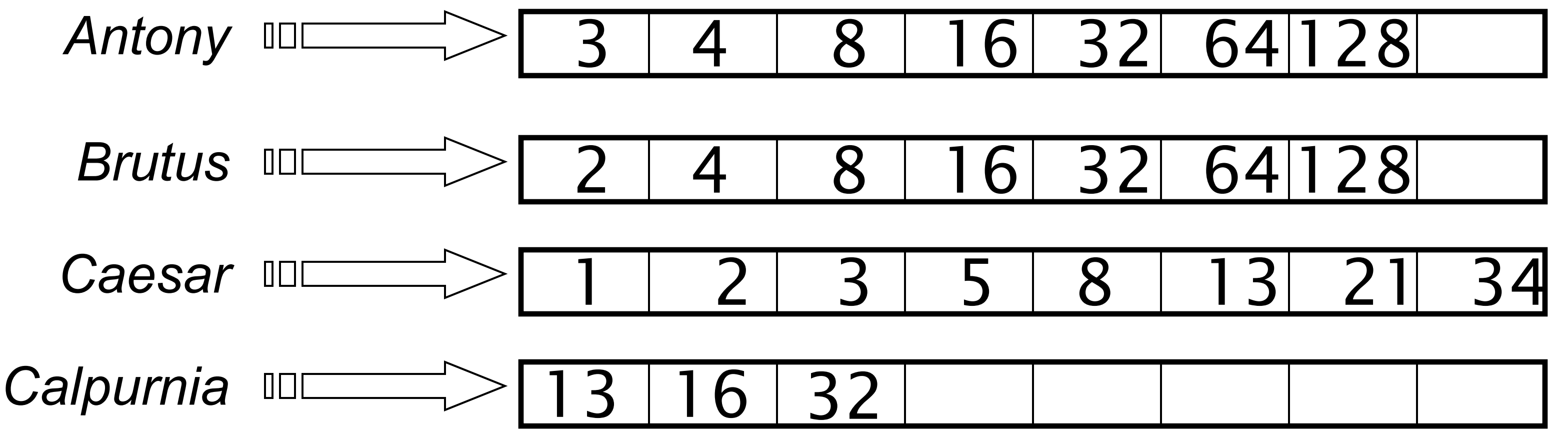


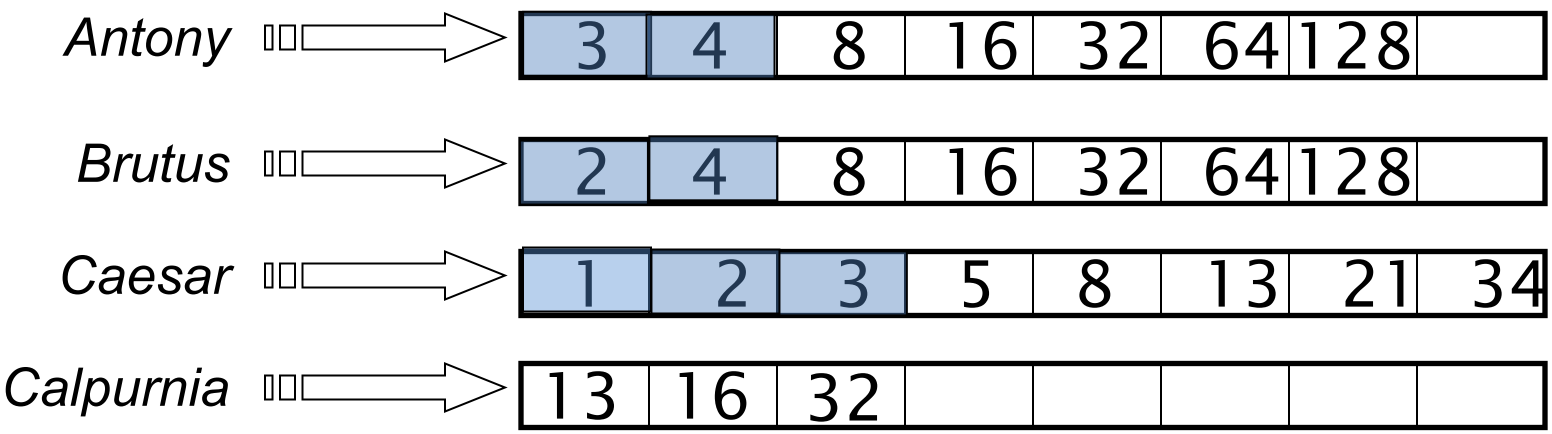
Image source: https://web.stanford.edu/class/cs276/19handouts/lecture2-intro-boolean-1per.pdf

Retrieval process using inverted index

1. Fetch the posting lists of query terms
2. Traverse through posting lists, and calculate the relevance score for each document in the posting lists
3. Retrieve top *n* documents with the highest relevance scores



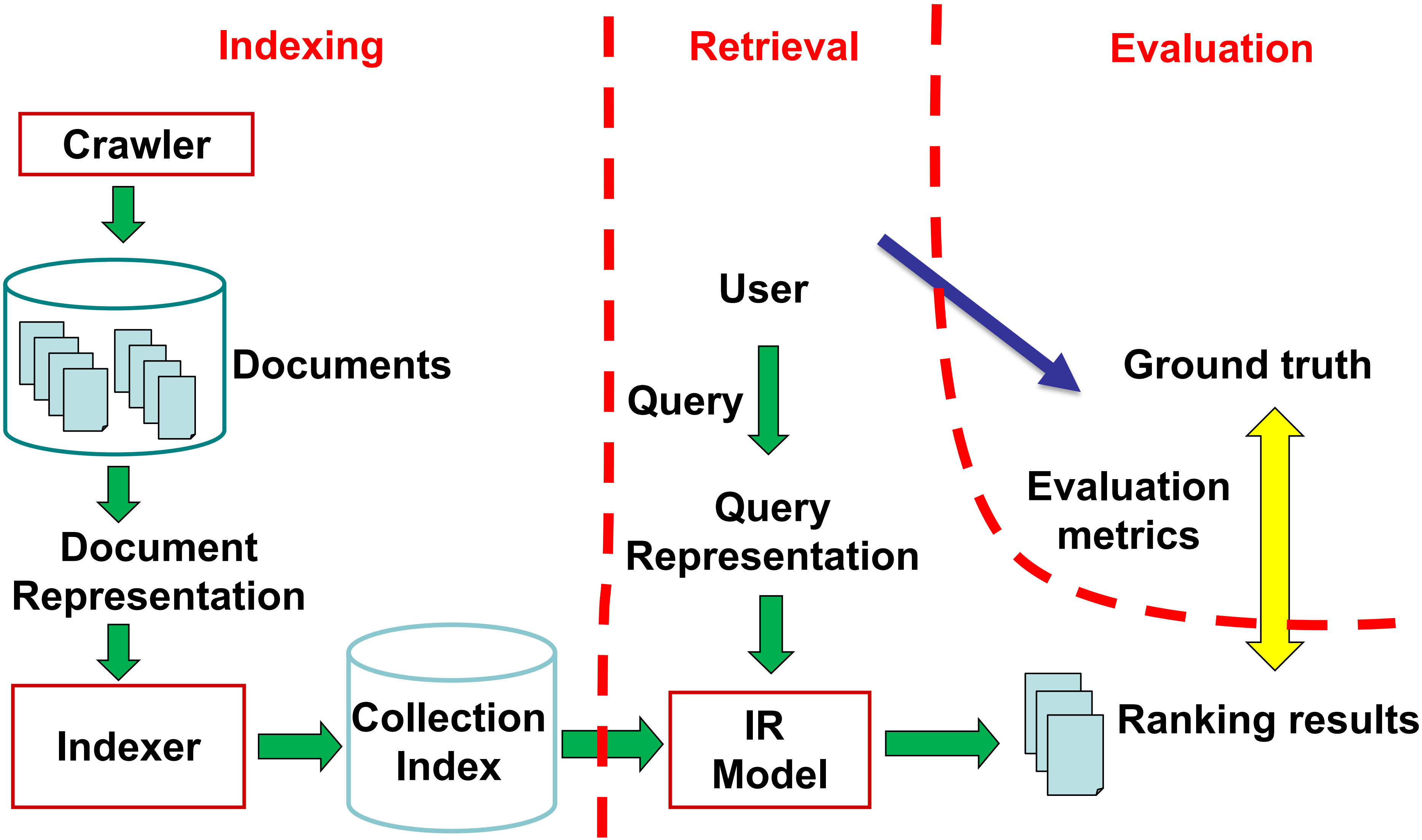
Search with concurrent traversal



Agenda

- Principles of Information Retrieval
- **Evaluation of a ranked list**
- IR with deep learning

Components of an IR System (simplified)



IR evaluation

- Evaluation of an IR system requires three elements:
  - A benchmark document collection
  - A benchmark suite of queries
  - Relevance judgements for pairs of query–document
    - Judgements specifies whether the document addresses the underlying information need of the query
    - Ideally done by human, but also through user interactions
    - Relevance judgements appear in forms of …
      - Binary: 0 (non-relevant) vs. 1 (relevant), or …
      - Multi-grade: more nuanced relevance levels, e.g. 0 (non-relevant), 1 (fairly relevant), 2 (relevant), 3 (highly relevant)

Evaluation Campaigns

- Text REtrieval Conference (TREC)

https://trec.nist.gov

- Conference and Labs of the Evaluation Forum (CLEF)

http://www.clef-initiative.eu

- MediaEval Benchmarking Initiative for Multimedia Evaluation

http://www.multimediaeval.org

Sample relevance judgement – format in research!

- TREC QRel (QueryRelevance) file: standard format to provide relevance judgements of some queries regarding to some documents

```
qry_id (iter)    doc_id   relevance_grade

 101    0        183294       0
 101    0        123522       2
 101    0        421322       1
 101    0        12312        0
                …
 102    0        375678       2
 102    0        123121       0
                …
 135    0        124235       0
 135    0        425591       1
                …
```

Common IR Evaluation Metrics

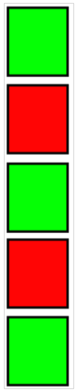
- Binary relevance
  - Precision@*n* (P@*n*)
  - Recall@*n* (P@*n*)
  - Mean Reciprocal Rank (MRR)
  - Mean Average Precision (MAP)

- Multi-grade relevance
  - Normalized Discounted Cumulative Gain (NDCG)

Precision@*n*

- **Precision@*n***: fraction of retrieved docs at top-*n* results that are relevant

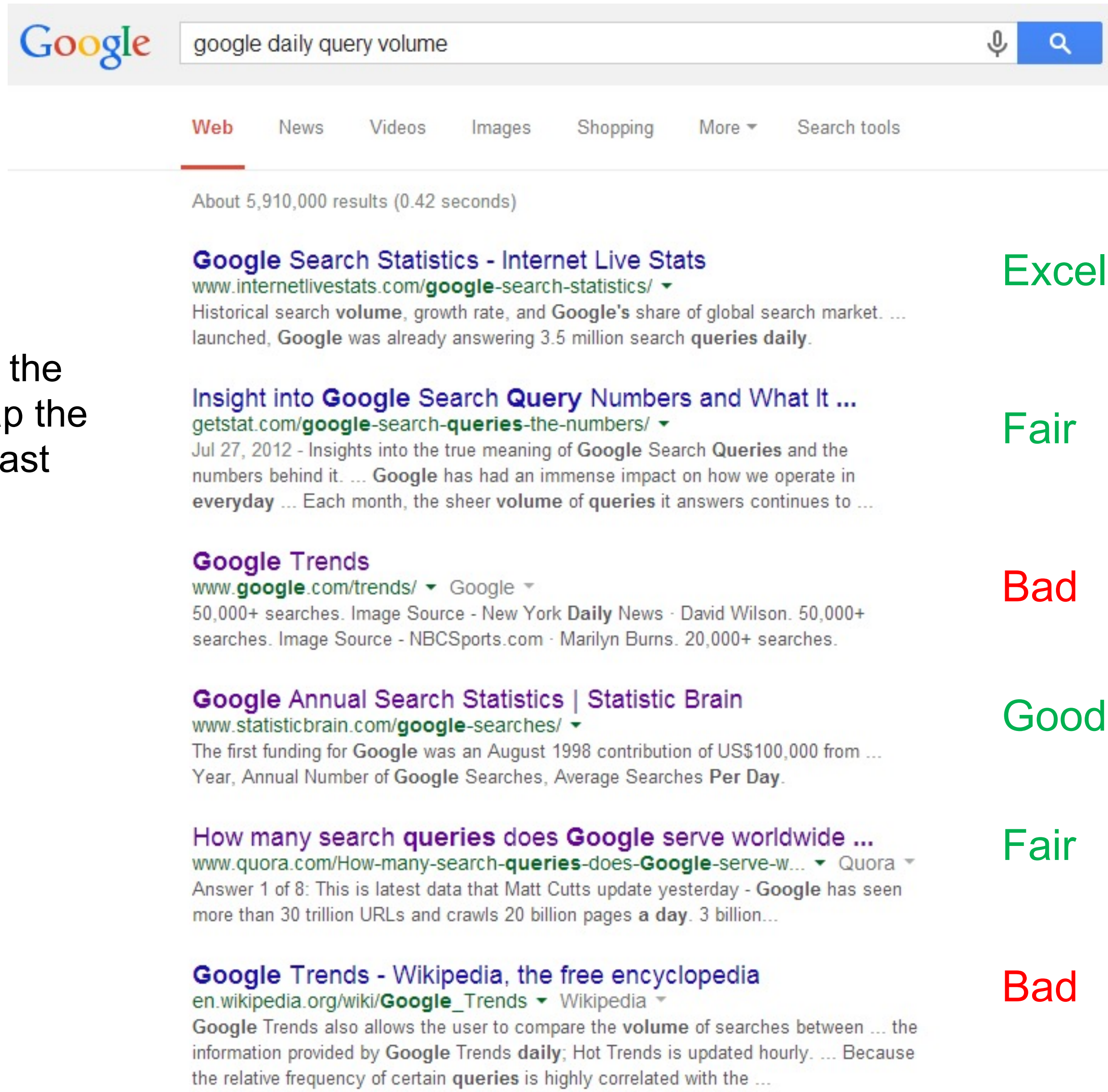
- Example:
  - P@3 = 2/3
  - P@4 = 2/4
  - P@5 = 3/5



- Final evaluation result is the mean of P@*n* across all queries in test set

Rank positions matter!

P@6 remains the same if we swap the first and the last result!



Excellent

Fair

Bad

Good

Fair

Bad

Discounted Cumulative Gain (DCG)

- A popular measure for evaluating web search and other related tasks

- Assumptions:
  - Highly relevant documents are more useful than marginally relevant documents (multi-grade relevance)
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined
    - This common behavior of users when interacting with ranked lists is known as *position bias*

Discounted Cumulative Gain (DCG)

- Gain: define gain as graded relevance, provided by relevance judgements

- Discounted Gain: gain is reduced as going down the ranking list. A common discount function: ${}^{1}\!/_{\log_2(\text{rank position})}$
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

- Discounted Cumulative Gain: the discounted gains are accumulated starting at the top of the ranking to the lower ranks till rank n

Discounted Cumulative Gain (DCG)

- Given the ranking results of a query, DCG at the position n of the ranking list is:

$$\mathrm{DCG@}n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}$$

where rel_i is the graded relevance (in relevance judgements) of the document at position i of the ranking results

- Alternative formulation (commonly used):

$$\mathrm{DCG@}n = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

DCG Example

Rank	Retrieved document ID	Gain (relevance)	Discounted gain	DCG
1	$d20$	3	3	3
2	$d243$	2	2/1=2	5
3	$d5$	3	3/1.59=1.89	6.89
4	$d310$	0	0	6.89
5	$d120$	0	0	6.89
6	$d960$	1	1/2.59=0.39	7.28
7	$d234$	2	2/2.81=0.71	7.99
8	$d9$	2	2/3=0.67	8.66
9	$d35$	3	3/3.17=0.95	9.61
10	$d1235$	0	0	9.61

$$\mathrm{DCG@10} = 9.61$$

Normalized DCG (NDCG)

- DCG results of different queries are not comparable,
  - Based on the relevance judgements of queries, the ranges of *good* and *bad* DCG results can be different between queries

- To normalize DCG at ranking position n:
  - For each query, estimate Ideal DCG (IDCG) which is the DCG for the ranking list, sorted by relevance judgements
  - Calculate NDCG by dividing DCG by IDCG

- Final NDCG@n is the mean across all test queries

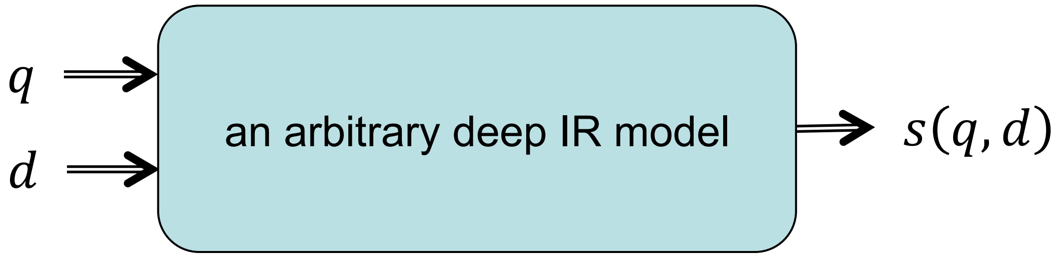
Agenda

- Principles of Information Retrieval
- Evaluation of a ranked list
- **IR with deep learning models**

Learning to predict relevance scores

- Instead of defining a formula as in classical IR models, we can learn to predict relevance scores $s(q, d)$ by training a neural network model

- Such neural/deep IR models can benefit from semantic relations in the embedding space, …
  - Hence do soft-matching between terms, in contrast to exact-matching in classical IR models



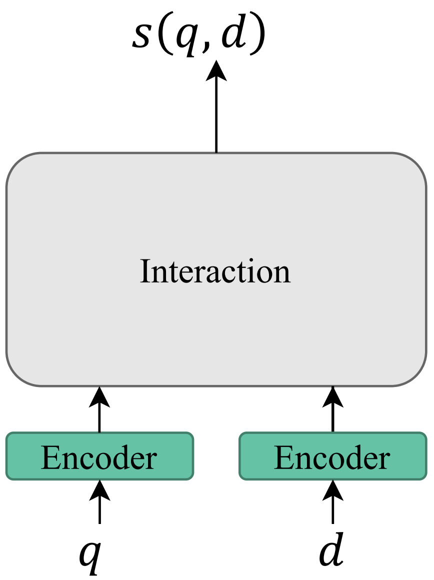
Elements of deep/neural IR models

- Interaction-based models
  - Model architecture
  - Training with Learning to Rank
  - Retrieval
- Dense Retrieval
  - Model architecture
  - Retrieval

Elements of deep/neural IR models

- **Interaction-based models**
  - **Model architecture**
  - **Training with Learning to Rank**
  - **Retrieval**
- Dense Retrieval
  - Model architecture
  - Retrieval

Model architecture



Interaction-based models

- calculate the *interactions* between the input embeddings of the document and query
- output a *feature vector*, representing the relation between query and document
- $s(q, d)$ is calculated from the feature vector

Task formulation

Training

- First, we want to learn a model that calculates the relevance score of a given query q to a given document d :

$$s(q, d)$$

Retrieval

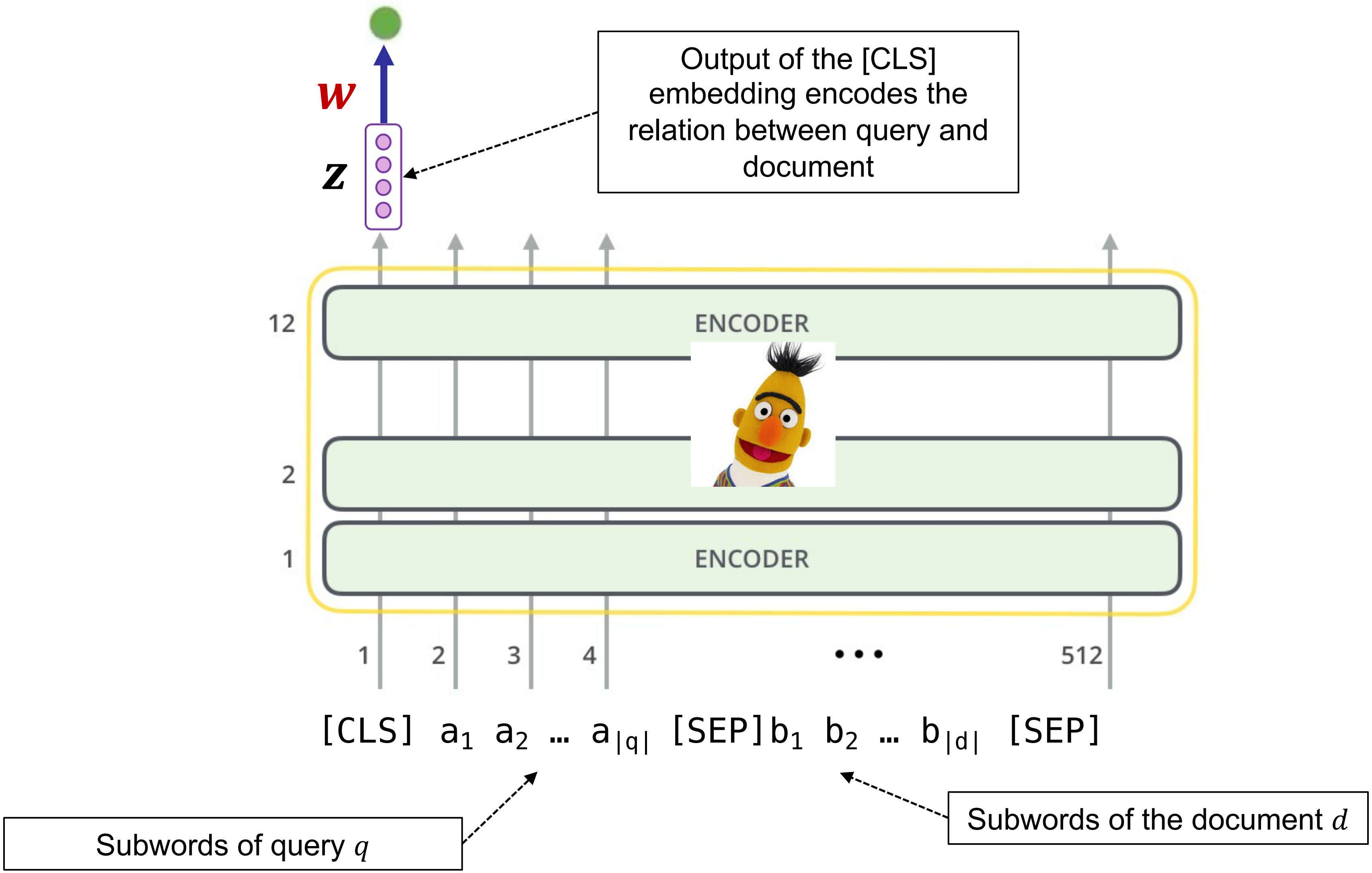
- for given q and the sets of documents $[d1, d2, d3, \dots, dM]$,
- the model calculates the relevance scores to all (or a subset) of documents:

$$[s(q, d1), s(q, d2), s(q, d3), \dots, s(q, dM)].$$

- This list is then sorted according to the predicted scores, and the top scoring documents are retrieved as results

BERT: an interaction-based IR model

$$s(q, d) = \boldsymbol{z}\boldsymbol{w}$$



Learning to Rank (LTR)

- It is insufficient to approach the learning of the models with ranking objectives, in the same way as the regression/classification models
- Consider the list of predicted scores by a model:

$$[s(q, d1), s(q, d2), s(q, d3), \dots, s(q, dM)]$$

- The final position of a document can only be known by comparing its predicted score with the ones of other documents
  - For example, only by looking at $s(q, d3) = 1.423$ we can not know in which position document $d3$ will end up

How should a model learn to predict scores according to a rank?!

- Learning to rank approaches:
  - Pointwise
  - Pairwise
  - Listwise (out of the scope of this lecture)

LTR – training data

- For a given query q, training data provides …
- a (small) set of relevant or positive documents:

$$[d_+^{(1)}, d_+^{(2)}, \dots]$$

  - d_+ is a document judged as relevant to q
  - For each query, usually only a few positive documents are available

- as well as a set of non-relevant or negative documents:

$$[d_-^{(1)}, d_-^{(2)}, \dots]$$

  - d_- can be a document judged as non-relevant to q, but also a randomly sampled document from the collection (why?)

Available collections with large training data

- MS MARCO (Microsoft MAchine Reading Comprehension)
  - Queries and retrieved passages of BING, annotated by human

	MS MARCO [28]
# of documents	8,841,822
Average document length	58.8 ± 23.5
Average query length	6.3 ± 2.6
# of training data points	39,780,811
# of validation queries	6,980
# of test queries	48,598

- TripClick (Collection & Log Files of a Health Web Search Engine)
  - Queries and clicked documents of TripDatabase search engine

Number of query-document interactions	4,054,593
Number of documents	1,523,878
Number of queries (all/HEAD/TORSO/TAIL)	692,699 / 5,879 / 108,314 / 578,506
Average query length	4.4 ± 2.4
Average document length	259.0 ± 81.7

https://microsoft.github.io/msmarco/
https://tripdatabase.github.io/tripclick/

Pointwise LTR

- Pointwise LTR models learn the relevance prediction of every positive/negative document <u>independently of the other documents</u>
  - Pointwise models are in fact classification/regression models

- Training data is therefore prepared in the form of:

`[input=(query, document), label(y)=relevance score]`

Example: For the query q

$$\left[\text{input} = \left(q, d_{+}^{(1)}\right), y = 1\right]$$
$$\left[\text{input} = \left(q, d_{+}^{(2)}\right), y = 1\right]$$
$$\left[\text{input} = \left(q, d_{+}^{(3)}\right), y = 1\right]$$
$$\dots$$
$$\left[\text{input} = \left(q, d_{-}^{(1)}\right), y = 0\right]$$
$$\left[\text{input} = \left(q, d_{-}^{(2)}\right), y = 0\right]$$
$$\left[\text{input} = \left(q, d_{-}^{(3)}\right), y = 0\right]$$
$$\dots$$

Pointwise LTR – loss

- Similar to classification tasks, Cross Entropy is a commonly used as the loss of pointwise LTR:

$$\mathcal{L} = -\mathbb{E}_{[(q,d),y]\sim\mathcal{T}}[y \log \sigma(s(q,d))]$$

  - $\mathcal{T} \rightarrow$ the set of training data
  - $\sigma(s(q,d)) \rightarrow$ sigmoid applied to the predicted score to turn the score into a probability

Pairwise LTR

- Pair-wise LTR is applied to pairs of positive-negative documents
- Pair-wise optimization aims to make the predicted score of a query to a relevant document higher than the one to a non-relevant document: $s(q, d_+) > s(q, d_-)$
  - This means that the IR model learns to give a higher relevance score to d_+ and therefore rank d_+ in a higher position than d_-. This (hopefully) leads to a better overall ranking results for the given query.
- The training data is therefore provided in the form of triplets:

```
[query, positive-document, negative-document]
```

Example: For the query q

$$
\begin{array}{ccc}
\left[q, d_+^{(1)}, d_-^{(1)}\right] & \left[q, d_+^{(2)}, d_-^{(1)}\right] & \dots \\
\left[q, d_+^{(1)}, d_-^{(2)}\right] & \left[q, d_+^{(2)}, d_-^{(2)}\right] & \dots \\
\left[q, d_+^{(1)}, d_-^{(3)}\right] & \left[q, d_+^{(2)}, d_-^{(3)}\right] & \dots \\
\dots & \dots & \dots
\end{array}
$$

Pairwise LTR – Max Margin loss

- Max-Margin is a widely used loss function for pair-wise training
  - Also called *Hinge loss*, *contrastive loss*, or *margin objective*

- Max-Margin ranking loss "punishes" the network until a given margin hyperparameter C is held between the predicted scores of the relevant and non-relevant documents:

$$\mathcal{L} = \mathbb{E}_{(q,d_+,d_-)\sim\mathcal{T}}[\max(0, C - (s(q,d_+) - s(q,d_-)))]$$

Examples when $C = 1$:

If $s(q,d_+) = 2$ and $s(q,d_-) = 1.8 \rightarrow \mathcal{L} = 0.8$

If $s(q,d_+) = 2$ and $s(q,d_-) = 3.8 \rightarrow \mathcal{L} = 2.8$

If $s(q,d_+) = 2$ and $s(q,d_-) = 0.8 \rightarrow \mathcal{L} = 0.0$

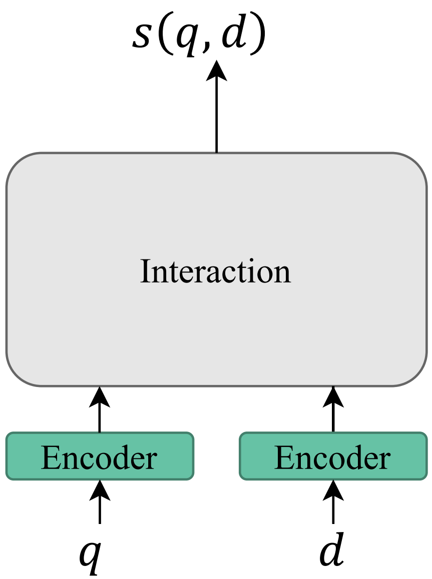
Retrieval with Interaction-based models

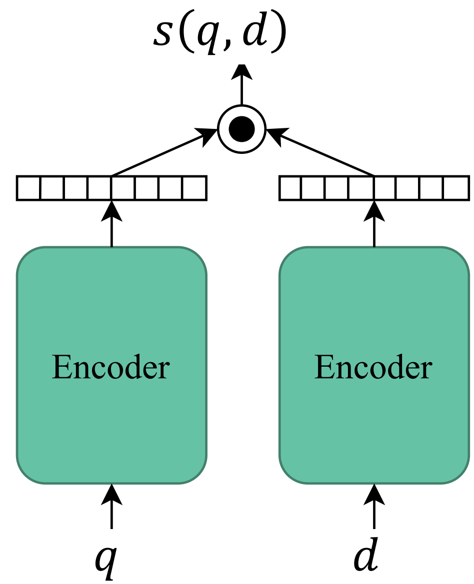
- Since neural/deep IR models are based on soft matching, they can’t benefit from inverted index for efficient retrieval
  - Neural/deep IR are therefore much slower in retrieval in comparison with exact-matching models
- Two (non-optimal) approaches:
  - Full-ranking: given a query, calculate relevance scores for all documents, sort the results, and retrieve the documents with highest relevance scores
    - Very expensive!
  - Re-ranking: re-rank top-t results of another IR model
    1. Pass the query to an efficient IR model (like BM25) and retrieve a first set of documents
    2. Select the top-t documents of this first set
    3. Re-calculate relevance scores for these documents using the neural IR model
    4. Update the original ranking results by re-ordering (re-ranking) the top-t documents based on the newly calculated scores

Elements of deep/neural IR models

- Interaction-based models
  - Model architecture
  - Training with Learning to Rank
  - Retrieval
- **Dense Retrieval**
  - **Model architecture**
  - **Retrieval**

Model architectures





Interaction-based models

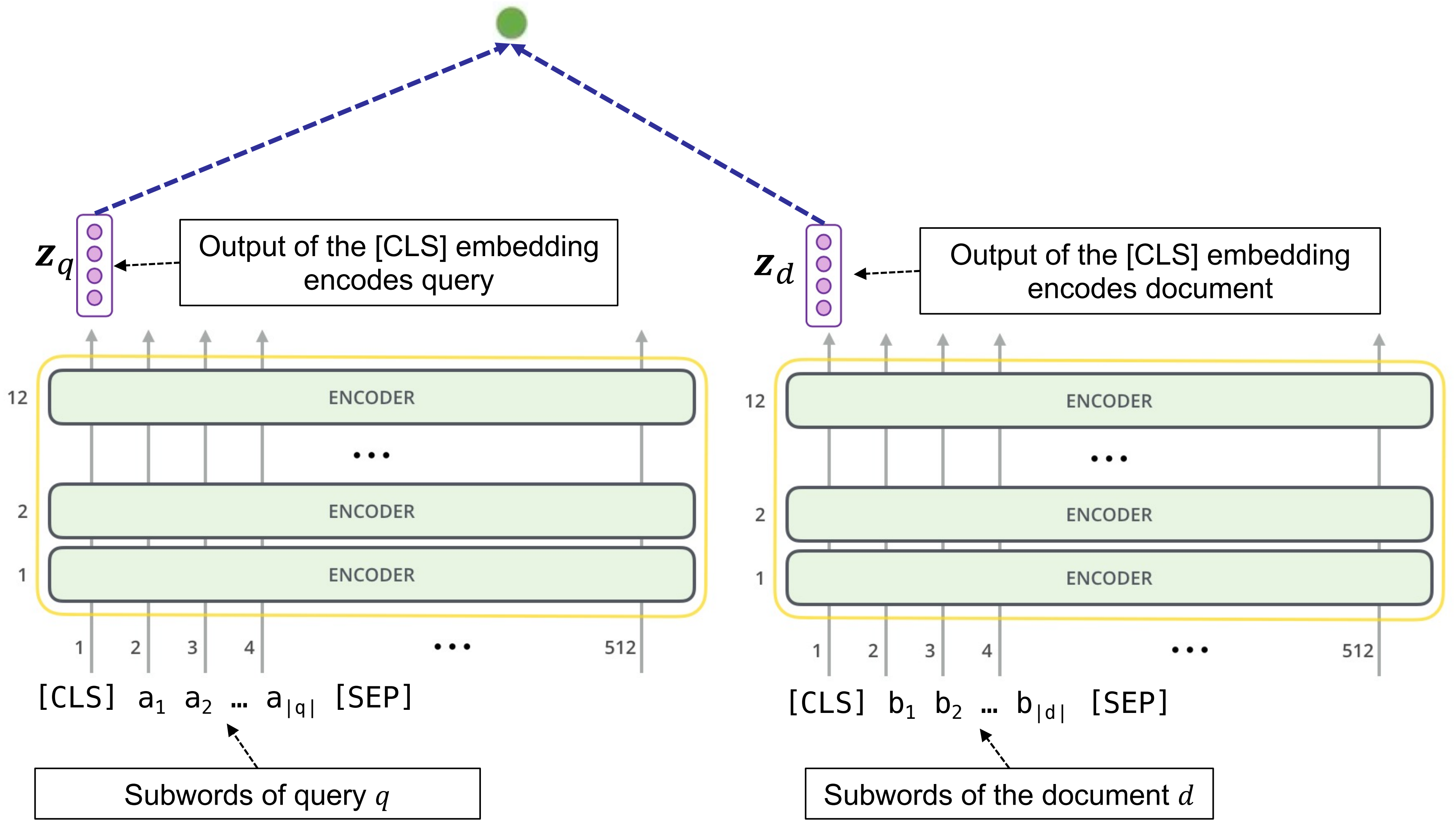
- calculate the *interactions* between the input embeddings of the document and query
- output a *feature vector*, representing the relation between query and document
- $s(q, d)$ is calculated from the feature vector

Dense Retrieval models

- first *encode* the document and the query in two separate vectors
- $s(q, d)$ is then calculated as the *similarity* of the two vectors
- This method enables direct retrieval of documents, achieved by finding the document embeddings which appear at the nearest proximity of the embedding of a query

A Dense Retrieval model using BERT

$$s(q,d) = \boldsymbol{z}_q \boldsymbol{z}_d$$



Retrieval with Dense Retrieval models

- The architecture of Dense Retrieval models enables direct retrieval instead of full-ranking or re-ranking

To retrieve the set of relevant documents …

- After training the model, the embeddings of all documents ($\boldsymbol{z}_d$) are calculated and stored
  - Sometimes the embeddings are stored in the data structure of an Approximate Nearest Neighbor (ANN) algorithms. This is referred to as document indexing.
- Now given query q, …
  - First the embedding of the query ($\boldsymbol{z}_q$) is calculated
  - Then, the most similar document embeddings to $\boldsymbol{z}_q$ are found and retrieved
    - This can be highly efficient especially when an ANN algorithm is used

- Dense Retrieval models enable highly efficient retrieval (even comparable with classical IR models), but commonly show a weaker performance in comparison with Interaction-based models