A Probabilistic Framework for Time-Sensitive Search

Dhruv Gupta & Klaus Berberich
{dhgupta, kberberi}@mpi-inf.mpg.de

June 9, 2016

❶ Motivation

❷ Building Blocks for Time-Sensitive Search

❸ Temporal Intent Disambiguation

❹ Temporally Diversified Retrieval

❺ Summary

Time-Sensitive Queries

Explicit Temporal Queries

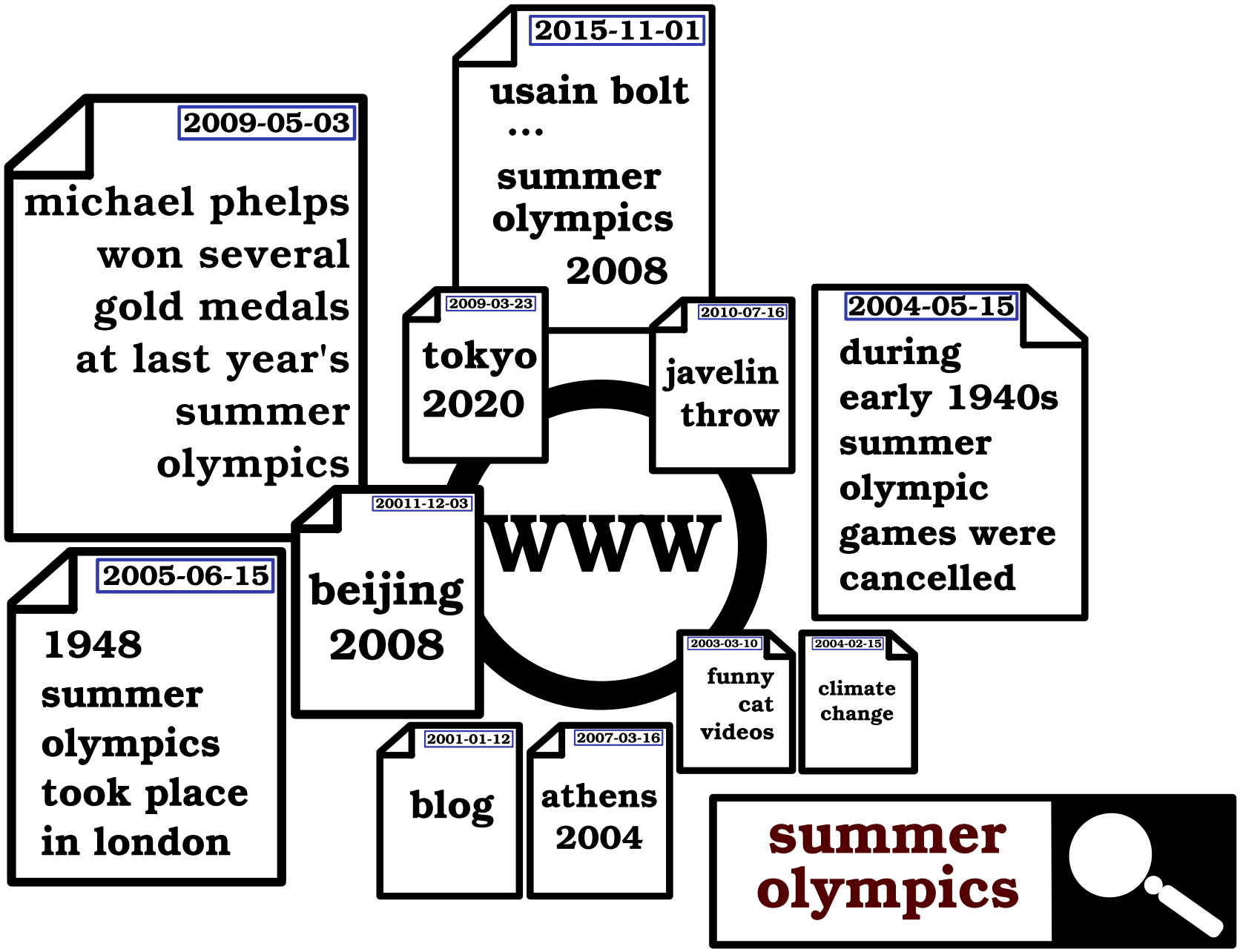
- 13.8% of Web queries [1]

Implicit Temporal Queries

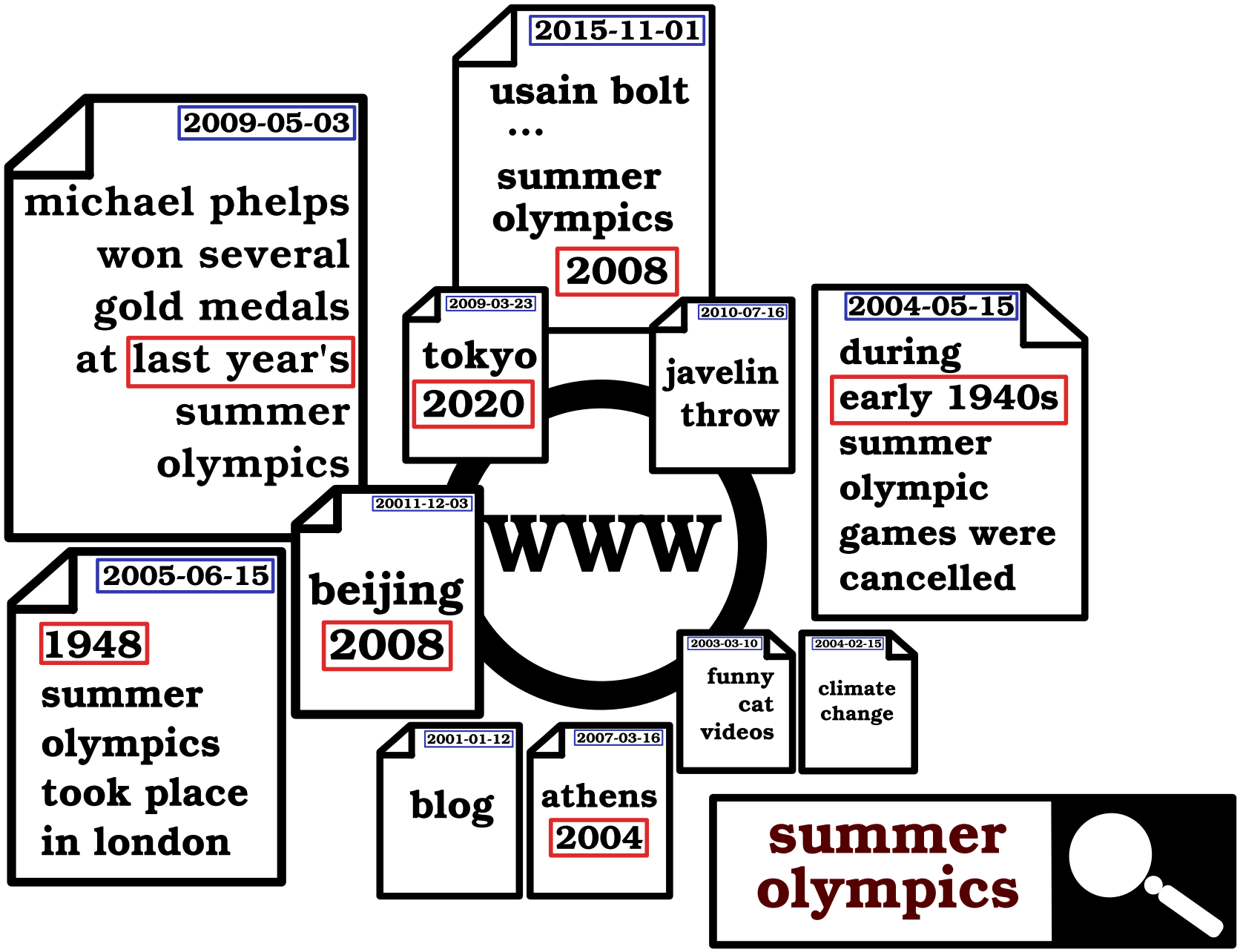
- 17.1% of Web queries [1]

[1] Kanahabua et al. : *Temporal Information Retrieval.* Foundations and Trends in Information Retrieval, 9(2):91-208, 2015.

Traditional Search



Time-Sensitive Search



1 Motivation

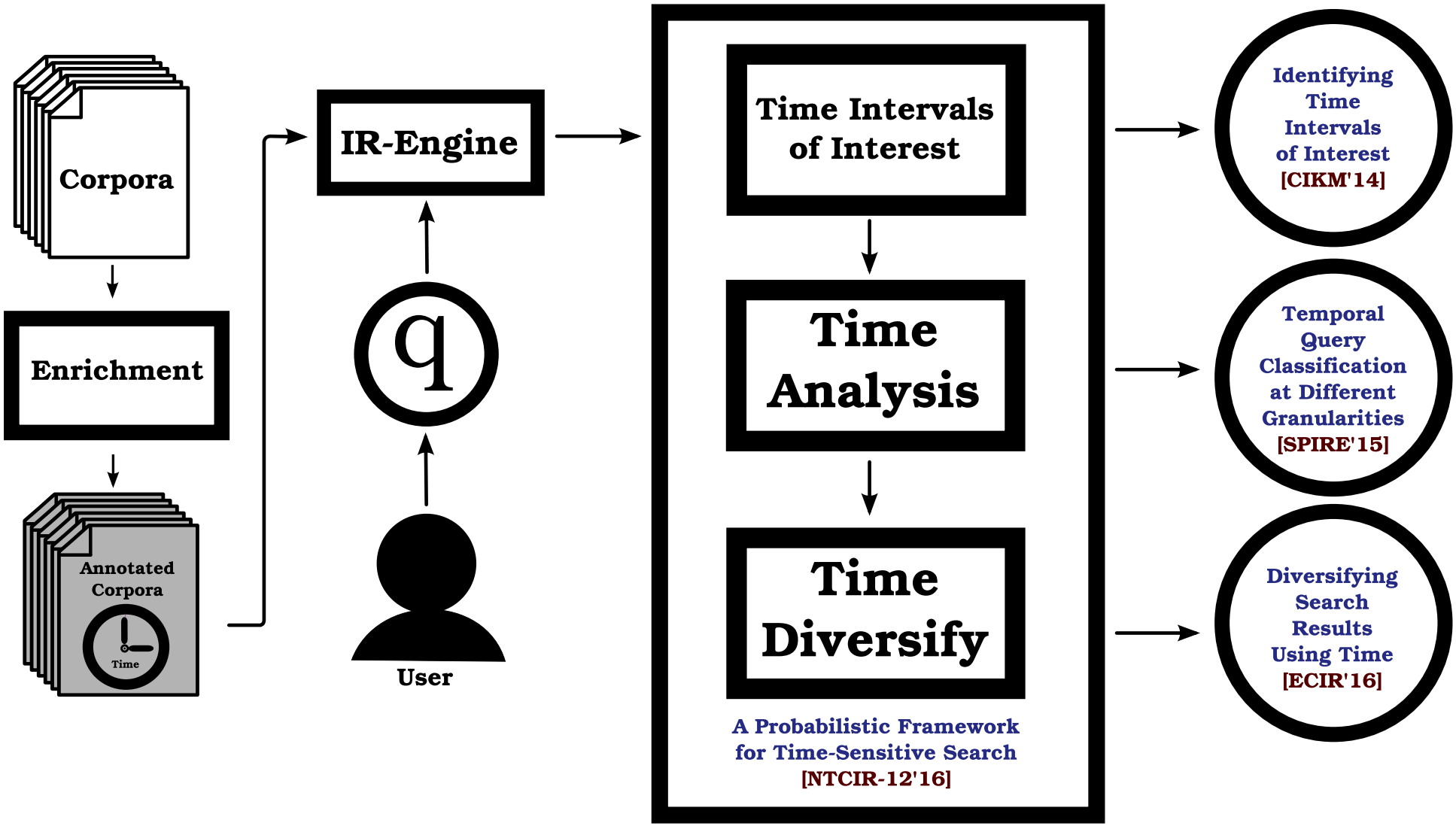
2 Building Blocks for Time-Sensitive Search

3 Temporal Intent Disambiguation

4 Temporally Diversified Retrieval

5 Summary

Building Blocks for Time-Sensitive Search



Time Model Incorporating Uncertainty [2]

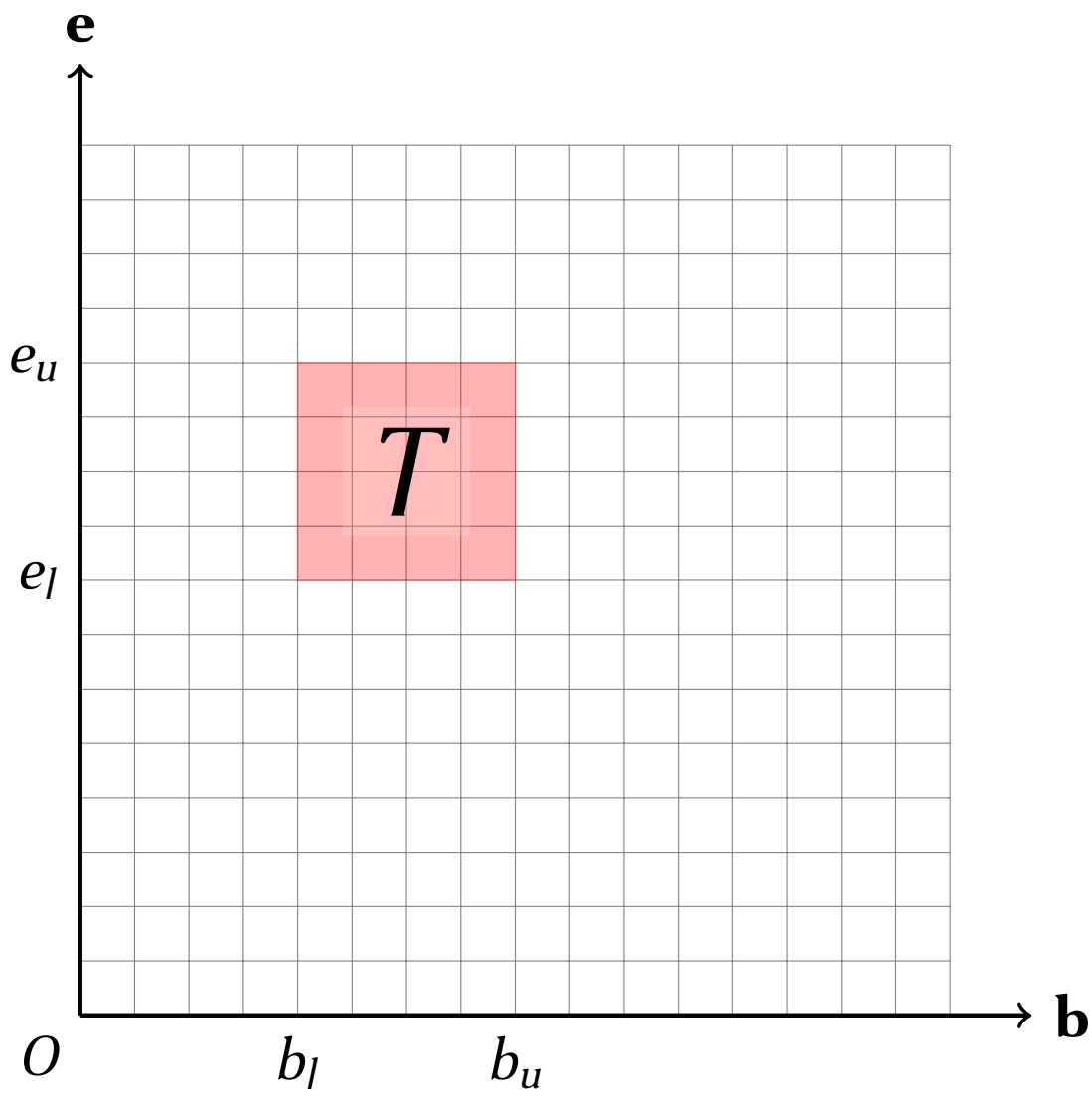
Time Model Incorporating Uncertainty

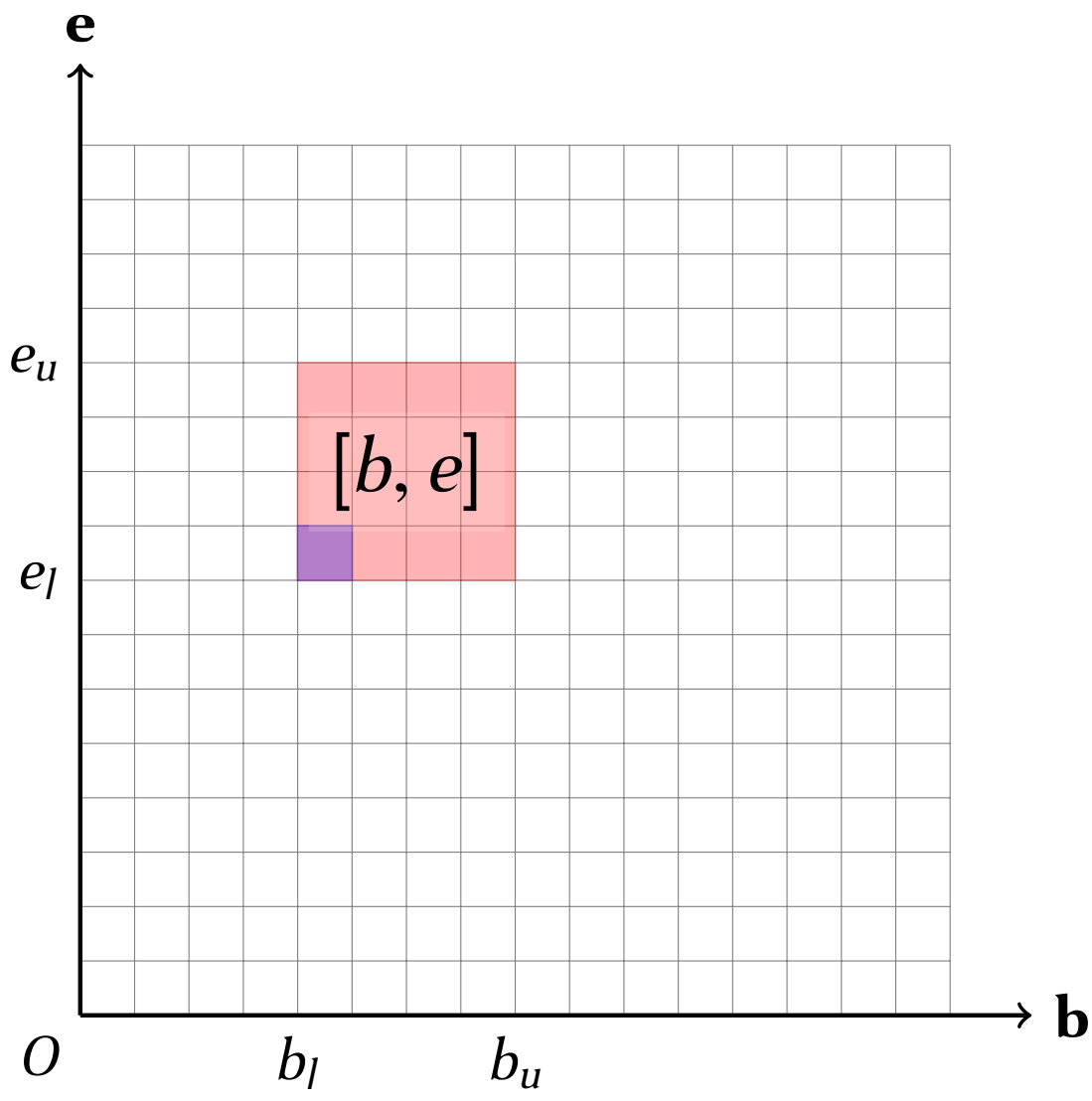
$$T = \langle b_l, b_u, e_l, e_u \rangle$$

Example

- Expression : *"1940s"*
- Resulting Temporal Expression (T) :
  $\langle 01-01-1940, 31-12-1949, 01-01-1940, 31-12-1949 \rangle$

[2] Berberich et al. : *A Language Modelling Approach for Temporal Information Needs.* ECIR 2010.


e
e_u
e_l
T
O
b_l
b_u
b


e
e_u
e_l
$[b, e]$
O
b_l
b_u
b

Identifying Interesting Time Intervals [3]

Hypothesis

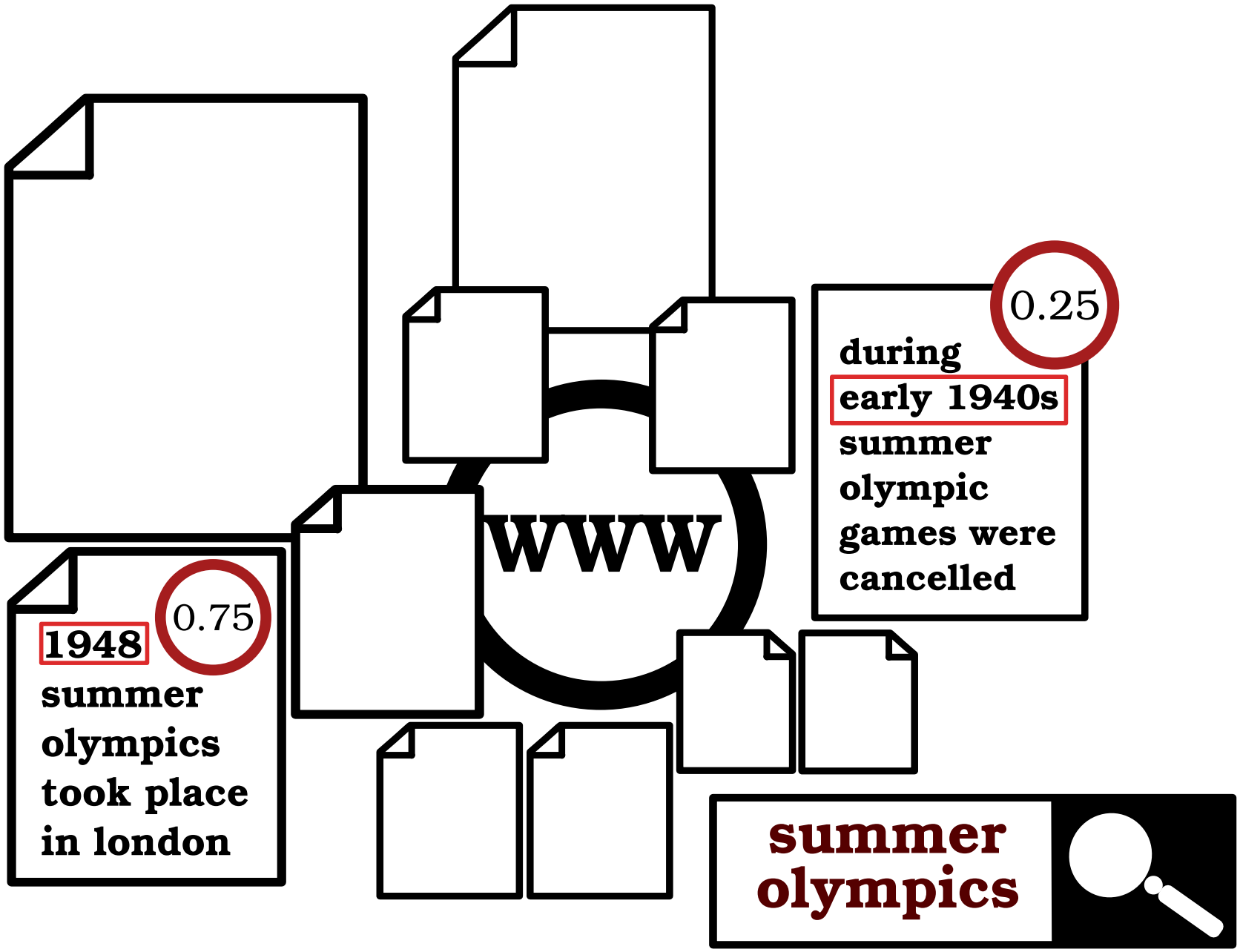
A time interval $[b, e]$ is interesting for a keyword query q, if it is frequently referred to by highly relevant documents.

Generative Model

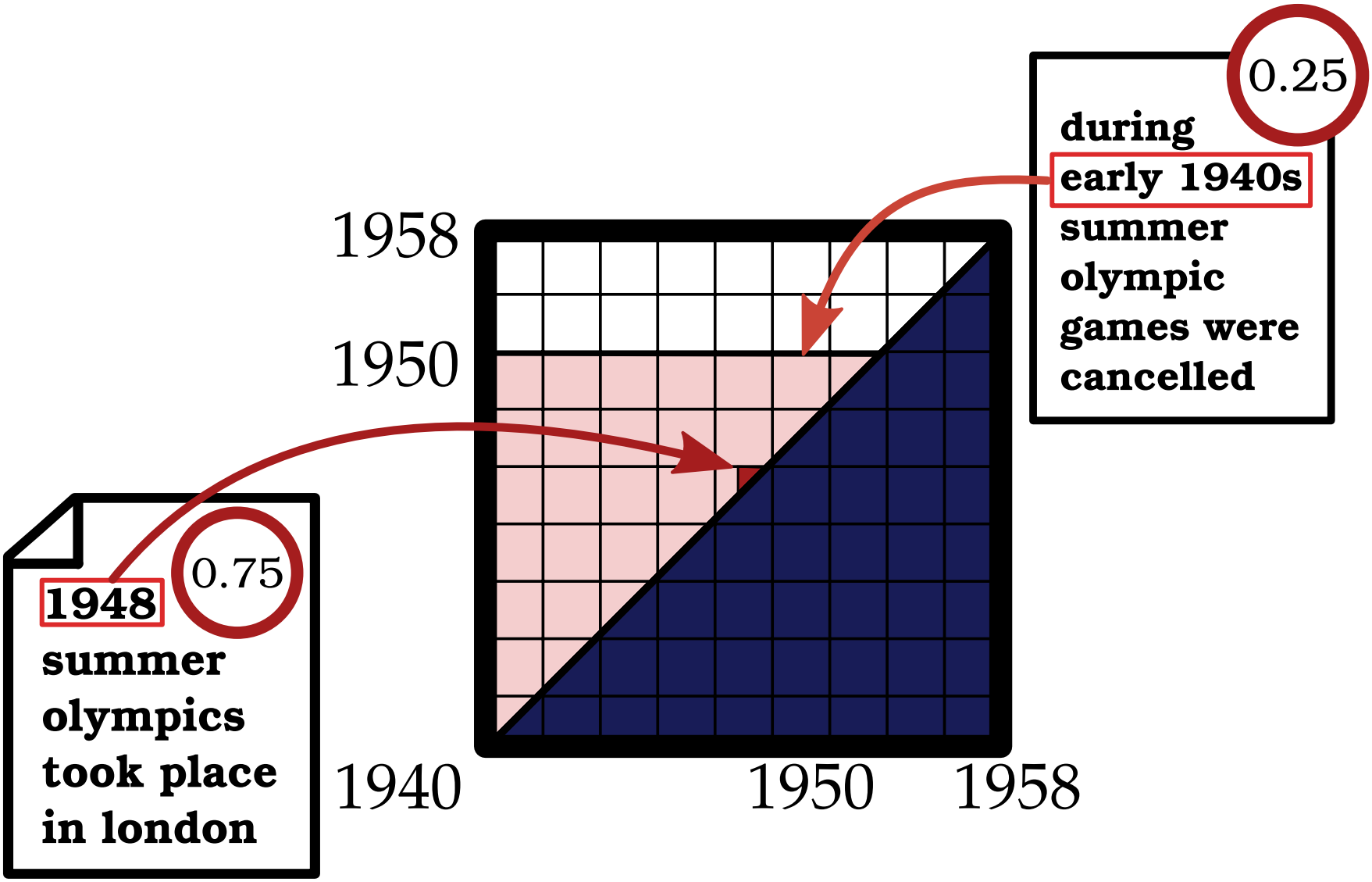
$$P([b, e]|q_{text}) = \sum_{d \in top(q,k)} P([b, e]|d_{time}) P(d|q_{text})$$

[3] Gupta & Berberich : *Identifying Time Intervals of Interest to Queries.* CIKM 2014.

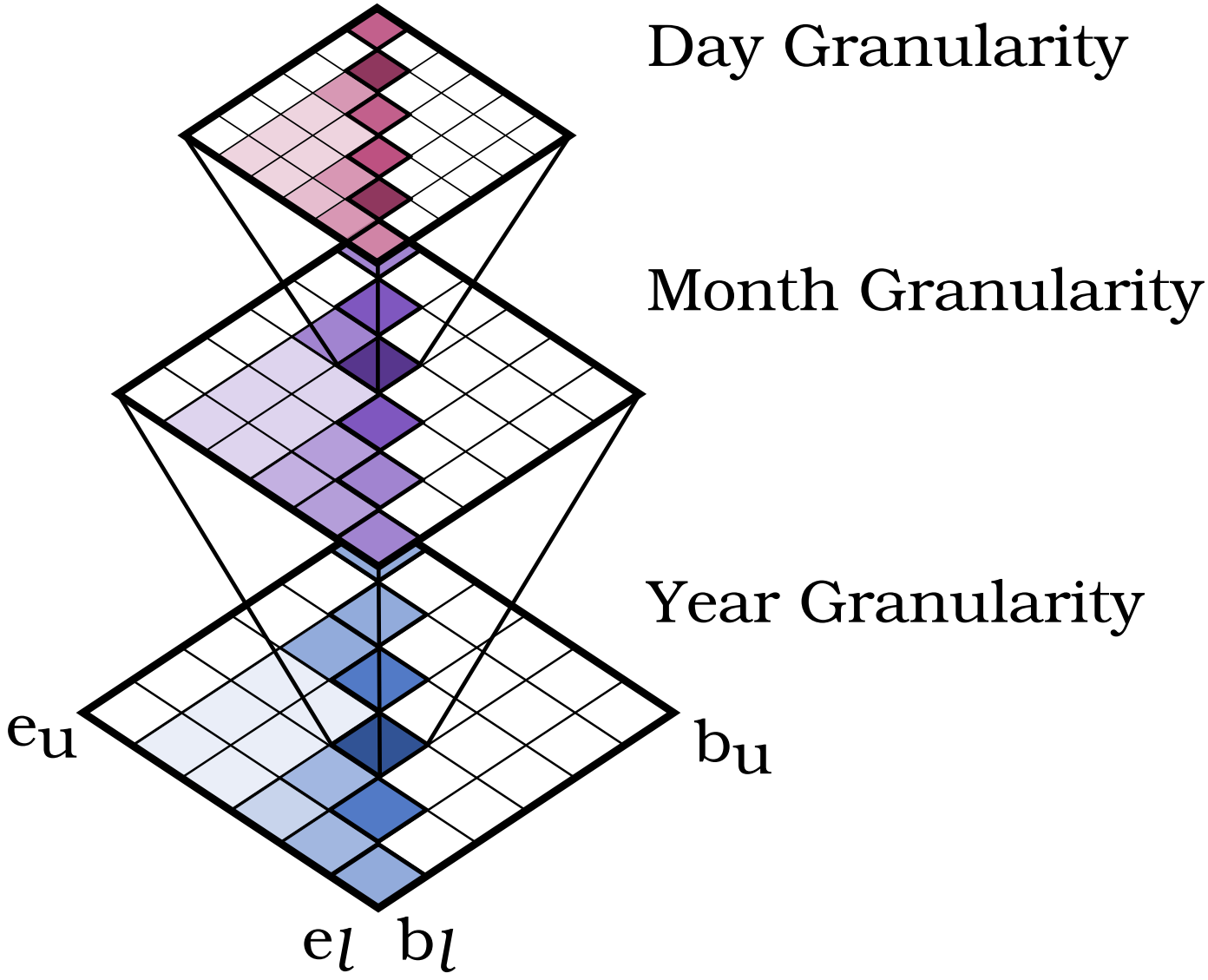
Counting Frequent Temporal Expressions



Counting Frequent Temporal Expressions



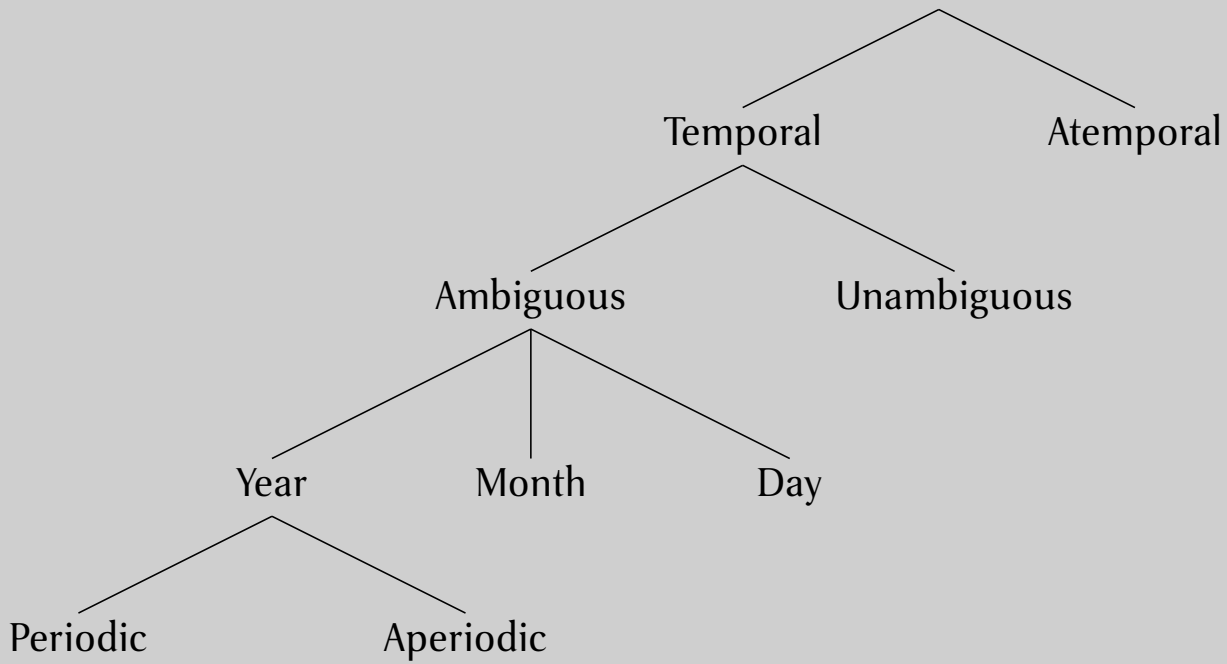
Counting Frequent Temporal Expressions Recursively



Identify Temporal Intents [4]

Contributions

- Identify temporal class in a taxonomy taking into account
  - Multiple granularities (day, month, year)
  - (A)periodicity of events
- Determine time intervals as intent for temporally ambiguous queries



[4] Gupta & Berberich : *Temporal Query Classification at Different Granularities. SPIRE* 2015.

Temporal Language Model [5]

Implicit Temporal Queries

- Query expansion of implicit temporal queries using interesting time intervals.

Temporal Language Model [5]

$$P(q|d) = P(q_{text}|d_{text}) \cdot P(q_{time}|d_{time})$$

$$P(q_{time}|d_{time}) = \prod_{[b,e] \in q_{time}} P([b,e]|d_{time})$$

[5] Berberich et al. : *A Language Modelling Approach for Temporal Information Needs.* ECIR 2010.

Diversifying Search Results Using Temporal Expressions[6]

- Retrospective overview of an entity or event
- Applications in digital humanities
- Search longitudinal document collections without knowledge of time intervals of interest

[6] Gupta & Berberich : *Diversifying Searach Results Using Time. ECIR* 2016.

[7] Photos from : `https://de.wikipedia.org/wiki/Mohandas_Karamchand_Gandhi`.

Diversify Search Results Using Temporal Expressions

- Adapt IA-Select [8] for diversification along time
- Query result set S that maximizes

$$\sum_{[b,e]\in q_{time}} P(\,[b,e] \mid q_{text}\,)\left(1-\prod_{d\in S}(1-P(\,q_{text} \mid d_{text}\,)P(\,[b,e] \mid d_{time}))\right)$$

[8] Agrawal et al. : *Diversifying Search Results. WSDM* 2009.

1 Motivation

2 Building Blocks for Time-Sensitive Search

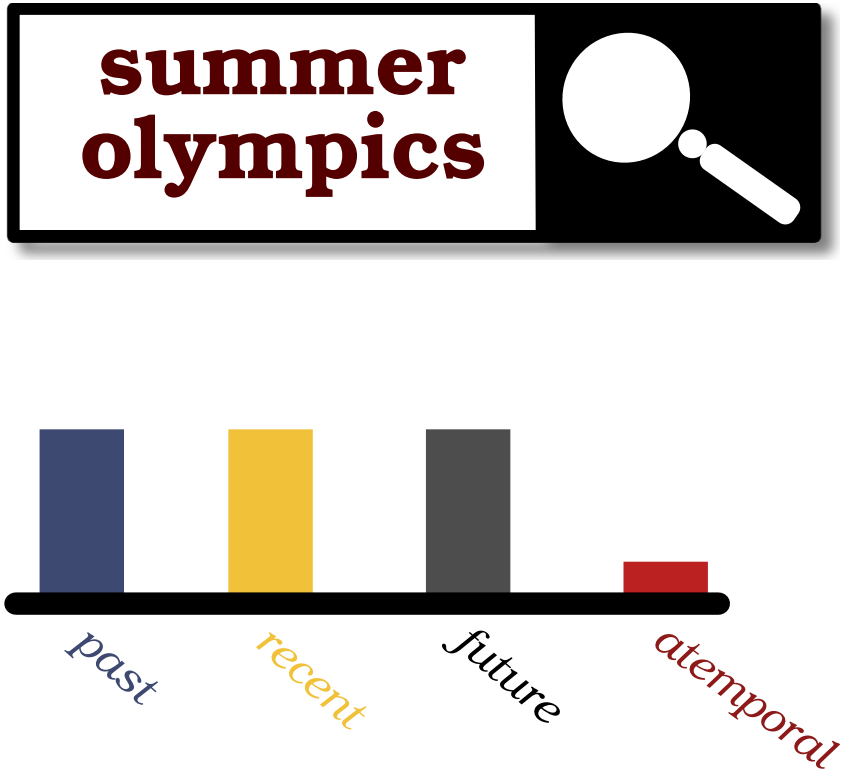
3 Temporal Intent Disambiguation

4 Temporally Diversified Retrieval

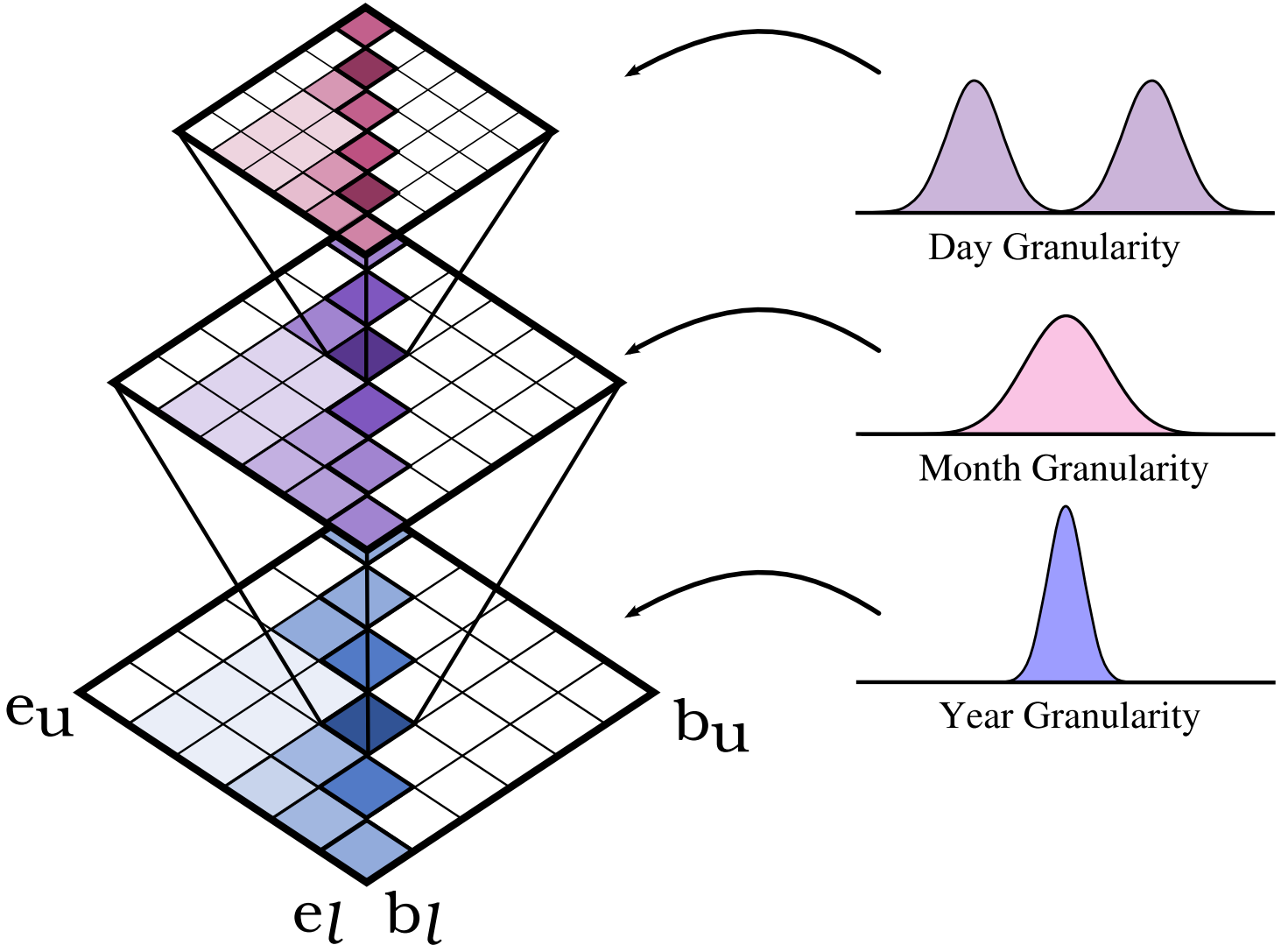
5 Summary

Problem Temporal Intent Disambiguation

- Given, a keyword query q_{text} and the classes C:
  - past
  - recent
  - future
  - atemporal
- Estimate $P(C|q)$



Approach — Analyze Time Intervals of Interest to Query



Approach — Analyze Time Intervals of Interest to Query

$$P(C = past|q) = \frac{1}{|\hat{q}_{time}|} \sum_{[b,e]\in\hat{q}_{time}} \mathbb{1}(t_{issue} > e)$$

$$P(C = recent|q) = \frac{1}{|\hat{q}_{time}|} \sum_{[b,e]\in\hat{q}_{time}} \mathbb{1}(b \leq t_{issue} \leq e)$$

$$P(C = future|q) = \frac{1}{|\hat{q}_{time}|} \sum_{[b,e]\in\hat{q}_{time}} \mathbb{1}(t_{issue} < b)$$

$$P(C = atemporal|q) = \sqrt{|\hat{q}_{time}|} \max_{[b,e]\in\hat{q}_{time}} |P([b,e]|q) - P([b,e]|D_{time})|$$

Results

System	Loss	Similarity	#Queries
Mpii-Tid-Formal	0.35	0.35	300
Mpii-Tid-Dry	0.34	0.39	20
Mpii-Tid-Train	0.30	0.48	73
Baseline	0.26	0.66	

Table: Results for our proposed system at different stages of the temporal intent disambiguation subtask.

Insights — *Good*

- Good results for following types of queries,
  i.e., low *loss* and high *similarity*:
  - `the advantages of hosting the olympic games`
  - `freedom of information act`
  - `when did ww2 start`
  - `how did bin laden die`
  - `when was television invented`
  - `history of slavery`
  - `occupy wall street movement`
- Insight: Queries that are history-oriented, i.e., have poignant past achieve good results

Insights — *Bad*

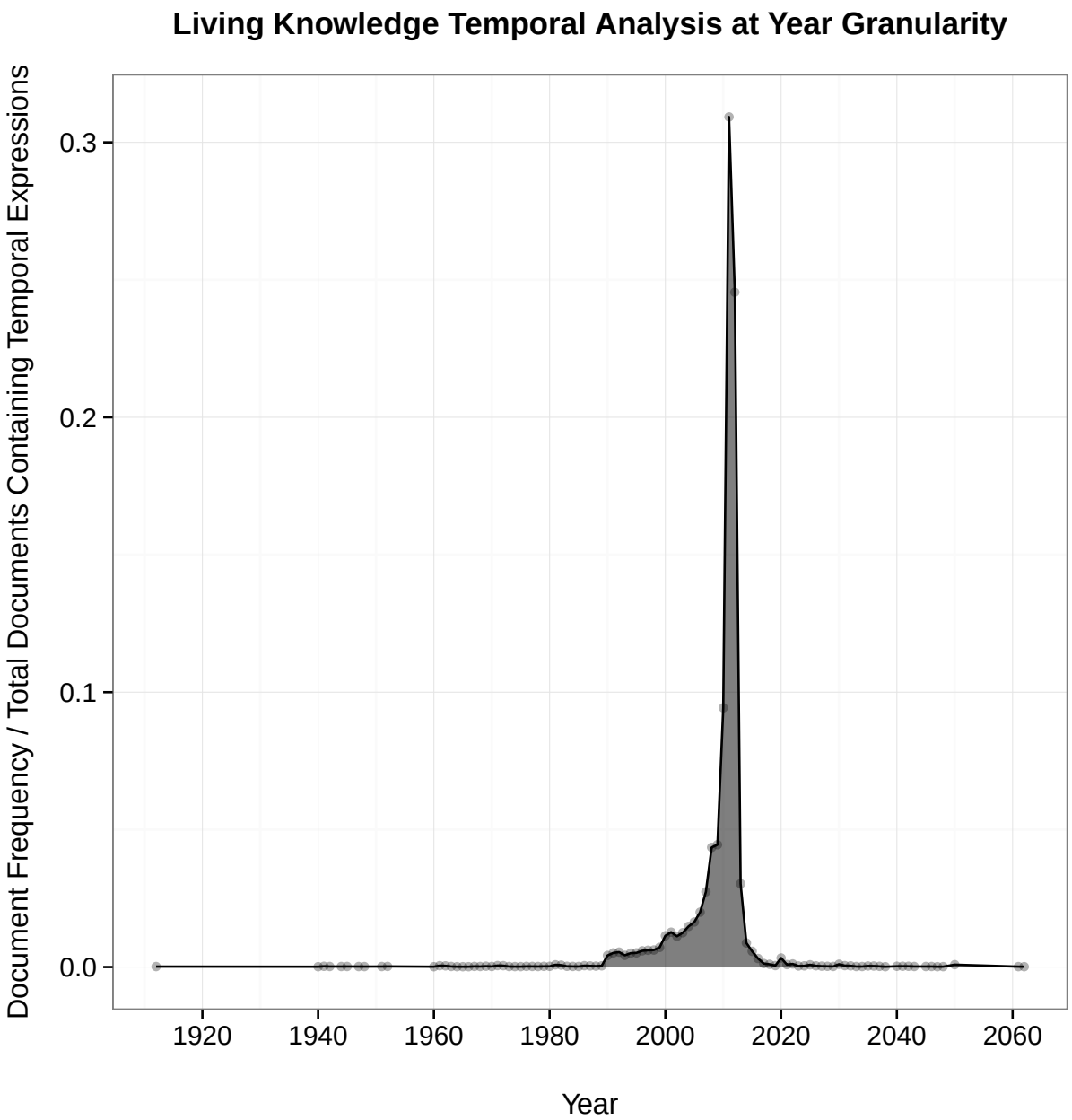
- Query examples with high *loss* and low *similarity*:
  - `naming university buildings with commercial brands`
  - `body posture alteration`
  - `dressing code in job interview`
  - `badminton games`
  - `advanced english`
  - `time warner austin`
- For these queries the interesting time intervals arose in [2011, 2013]

Insights — *Bad*

- Query examples with high *loss* and low *similarity*:
  - `naming university buildings with commercial brands`
  - `body posture alteration`
  - `dressing code in job interview`
  - `badminton games`
  - `advanced english`
  - `time warner austin`
- For these queries the interesting time intervals arose in [2011, 2013]

Why?

Insights — *Ugly*



1 Motivation

2 Building Blocks for Time-Sensitive Search

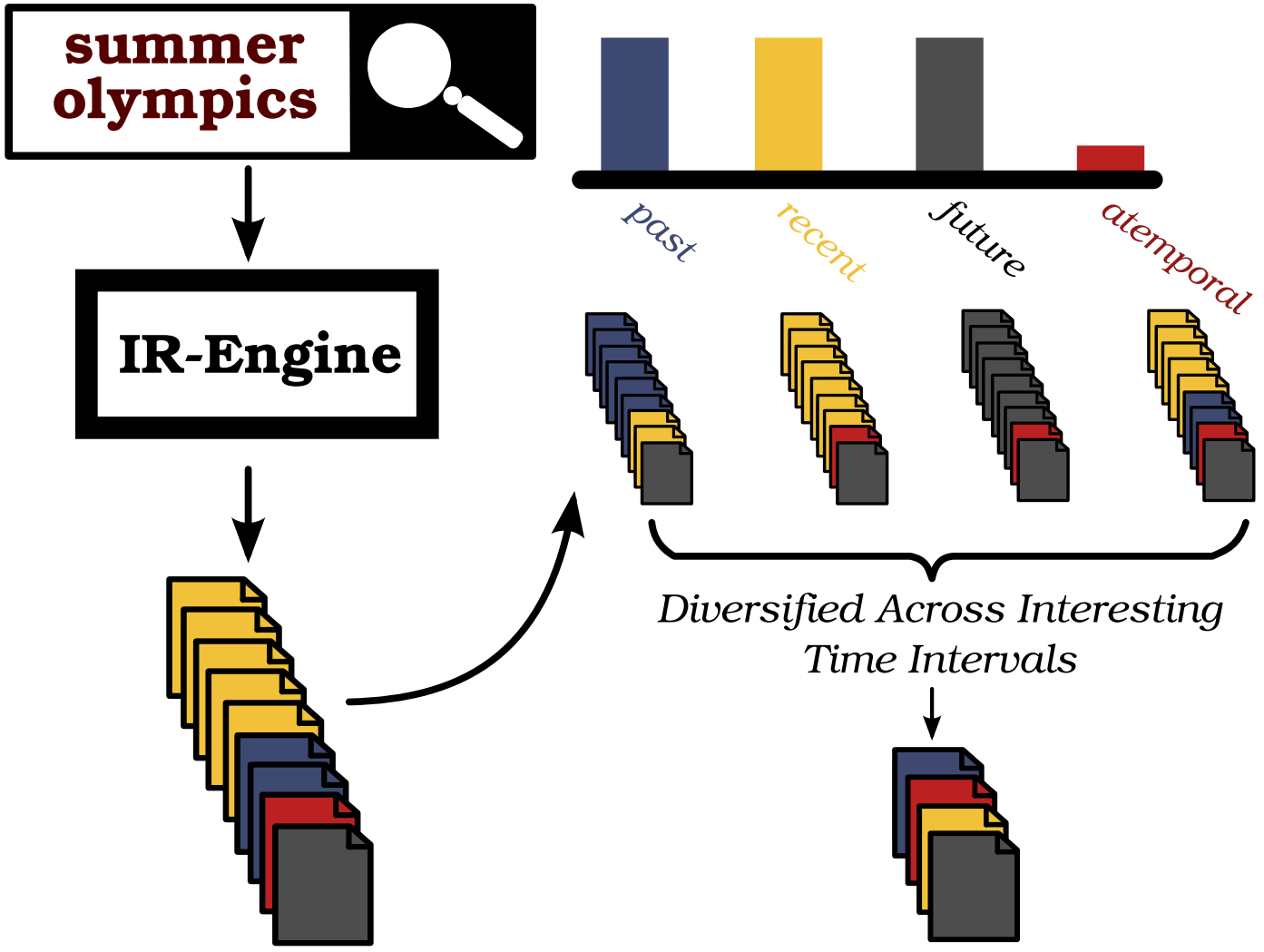
3 Temporal Intent Disambiguation

4 Temporally Diversified Retrieval

5 Summary

Problem Temporal Diversified Retrieval

Given, keyword query q_{text} and document collection D, estimate $P(d|q, C)$.



Approach

- Use the temporal language model to re-rank documents
- For $C = \textit{recent}$
  - Expand query with query issue time
- For $C = \textit{past}$
  - Expand query with time intervals that lie before query issue time
- For $C = \textit{future}$
  - Expand query with time intervals that lie after query issue time
- For $C = \textit{atemporal}$
  - Use the pseudo-relevant set of documents.
- For *diversified* set of documents
  - Use temporal diversification to find a set of documents such that the user sees at least one document from each of the interesting time intervals

Results — per Category Retrieval

Category	Dry-run nDCG@20	Formal-run nDCG@20
Atemporal	0.17	0.34
Past	0.19	0.39
Recent	0.05	0.34
Future	0.02	0.34
All	0.11	0.35

Table: Results for our proposed system for retrieving time-sensitive documents at different stages of the temporally diversified retrieval subtask.

Results — Temporal Diversification

Stage	**nDCG@20**	**D#-nDCG@20**
Dry-run	0.18	0.41
Formal-run	0.33	0.57

Table: Results for our proposed system for diversifying time-sensitive documents at different stages of the temporally diversified retrieval subtask.

Insights

- Overall comparing to organizers' system our method did not fare as well

Insights

- Overall comparing to organizers' system our method did not fare as well
- Why?
  - The role of the retrieval method for producing an initial set of pseudo-relevant documents
  - The role that document content temporal expressions play in our approach — we used annotations provided with corpus

Insights

- Overall comparing to organizers' system our method did not fare as well
- Why?
  - The role of the retrieval method for producing an initial set of pseudo-relevant documents
  - The role that document content temporal expressions play in our approach — we used annotations provided with corpus
- Improvements
  - Try different initial retrieval methods
  - Use an external temporal tagger (e.g., SuTime, HeidelTime) as opposed to temporal expressions provided with document collection

1 Motivation

2 Building Blocks for Time-Sensitive Search

3 Temporal Intent Disambiguation

4 Temporally Diversified Retrieval

5 Summary

Summary — Building Blocks for Time-Sensitive Search