

# Chapter IR:IV

## IV. Retrieval Models

- ❑ Overview of Retrieval Models
- ❑ Boolean Retrieval
- ❑ Vector Space Model
- ❑ Binary Independence Model
- ❑ Okapi BM25
- ❑ Divergence From Randomness
- ❑ Latent Semantic Indexing
- ❑ Explicit Semantic Analysis
- ❑ Language Models
- ❑ Combining Evidence
- ❑ Learning to Rank

# Overview of Retrieval Models

## Document Views

Information retrieval requires modeling and representing documents on a computer. We distinguish three **orthogonal** views on a document's content:

### 1. Layout view

Presentation of a document on a (two-dimensional) medium.

### 2. Structural / logical view

Composition and logical structure of a document. Example:

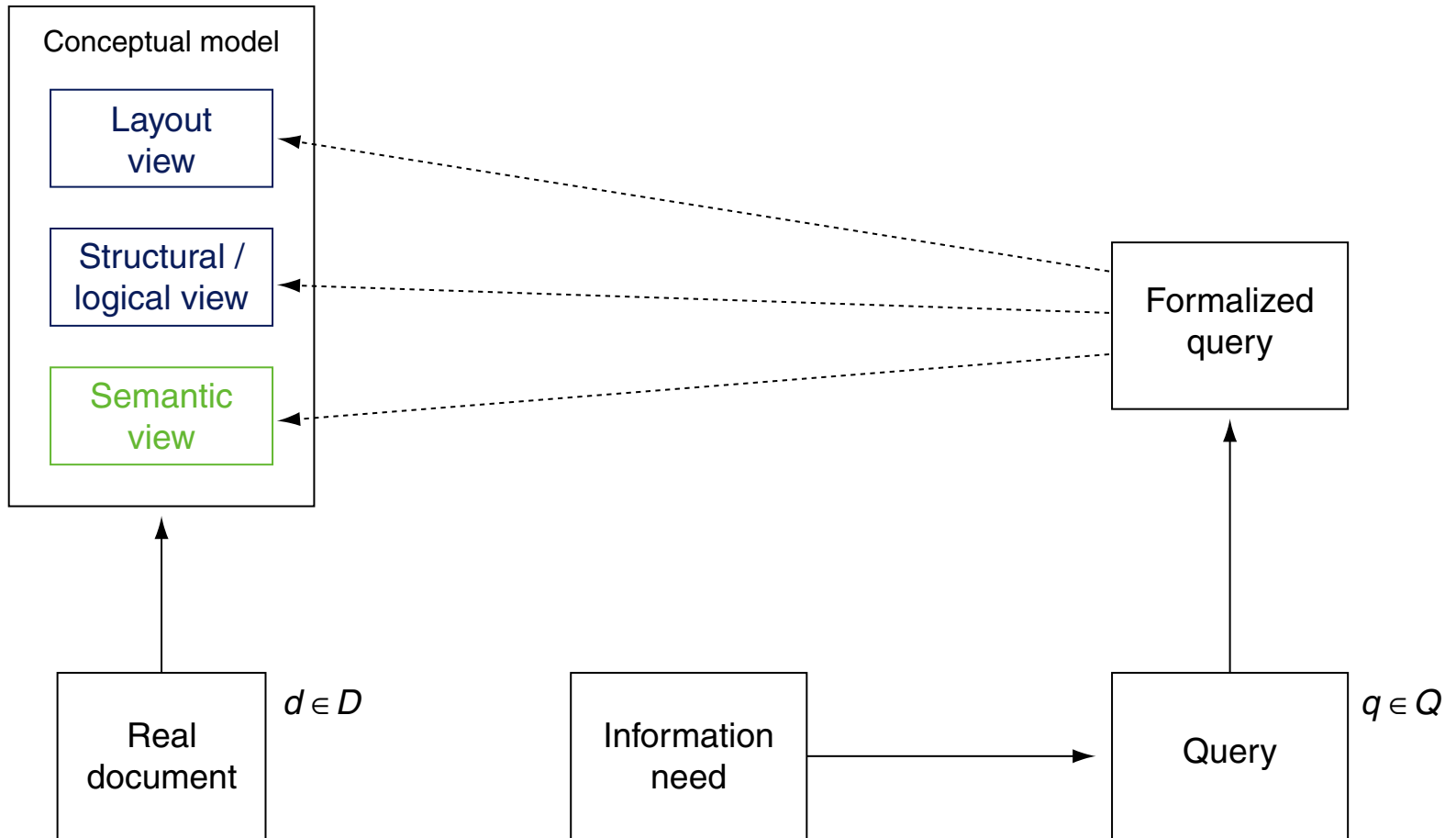
```
\documentclass[twocolumn,german]{article}  
\title{...}  
\author{...}  
\section{...}
```

### 3. Semantic view

The meaning of a document or its message, allowing for its interpretation.

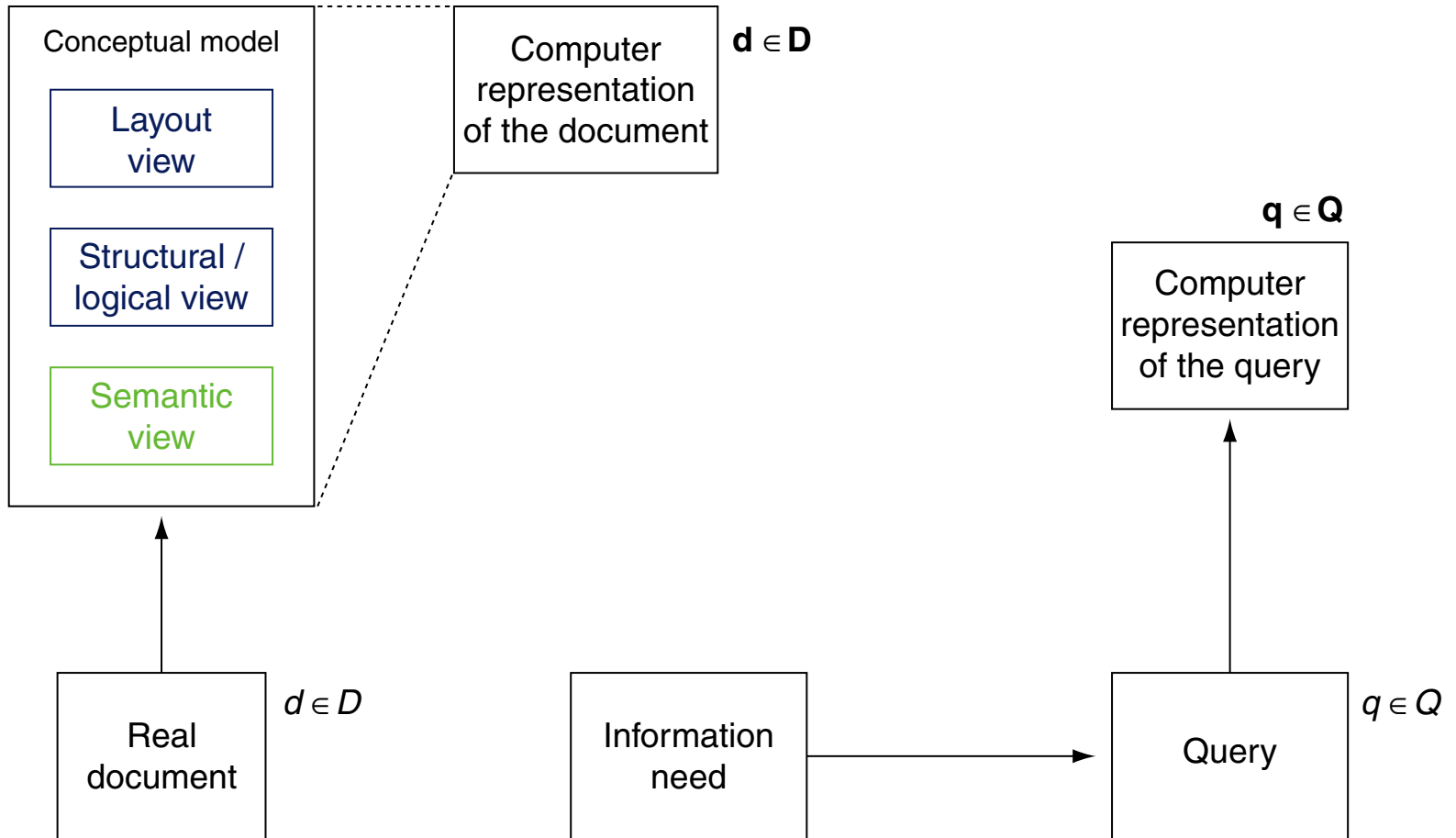
# Overview of Retrieval Models

## Retrieval Models



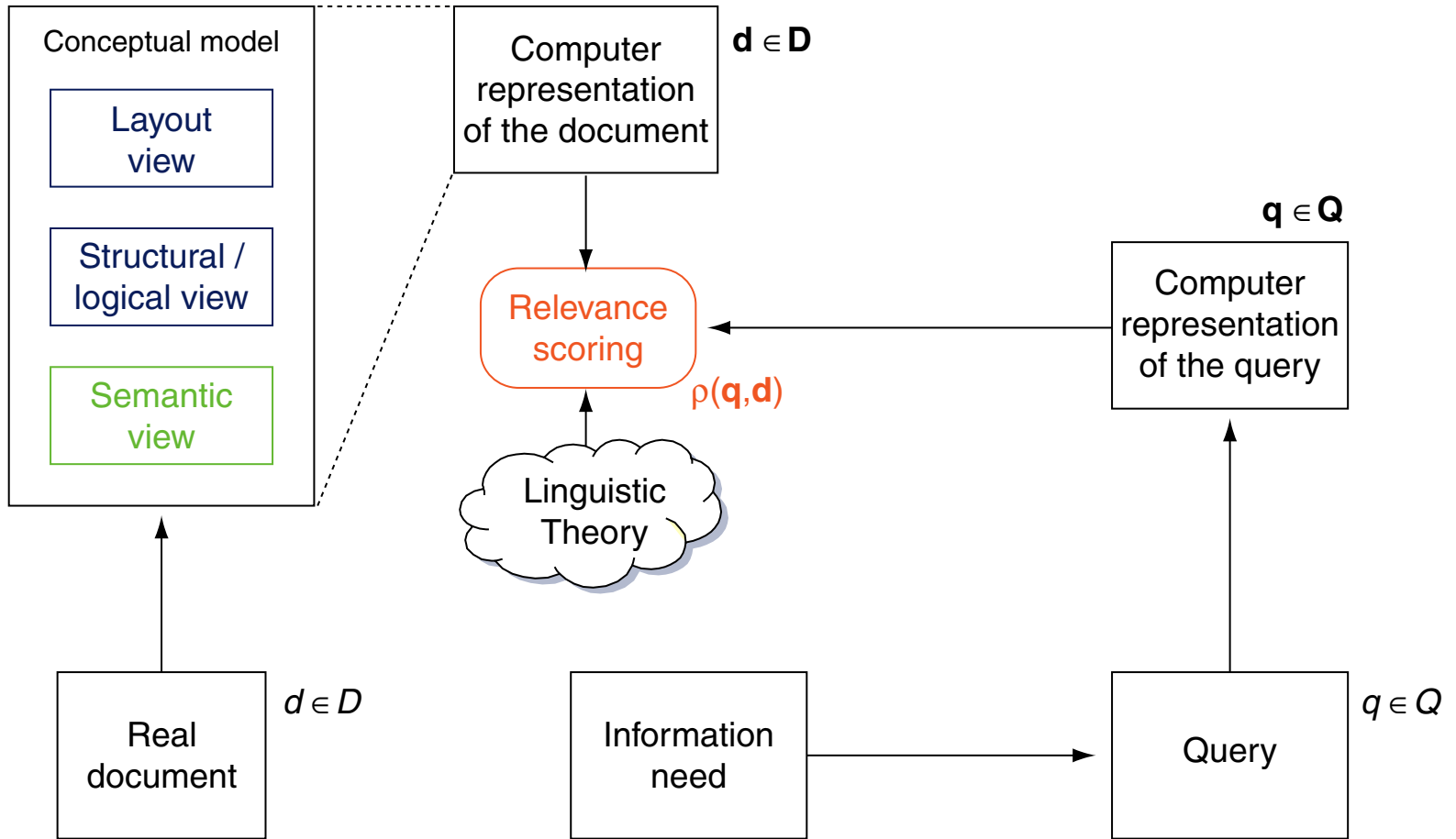
# Overview of Retrieval Models

## Retrieval Models



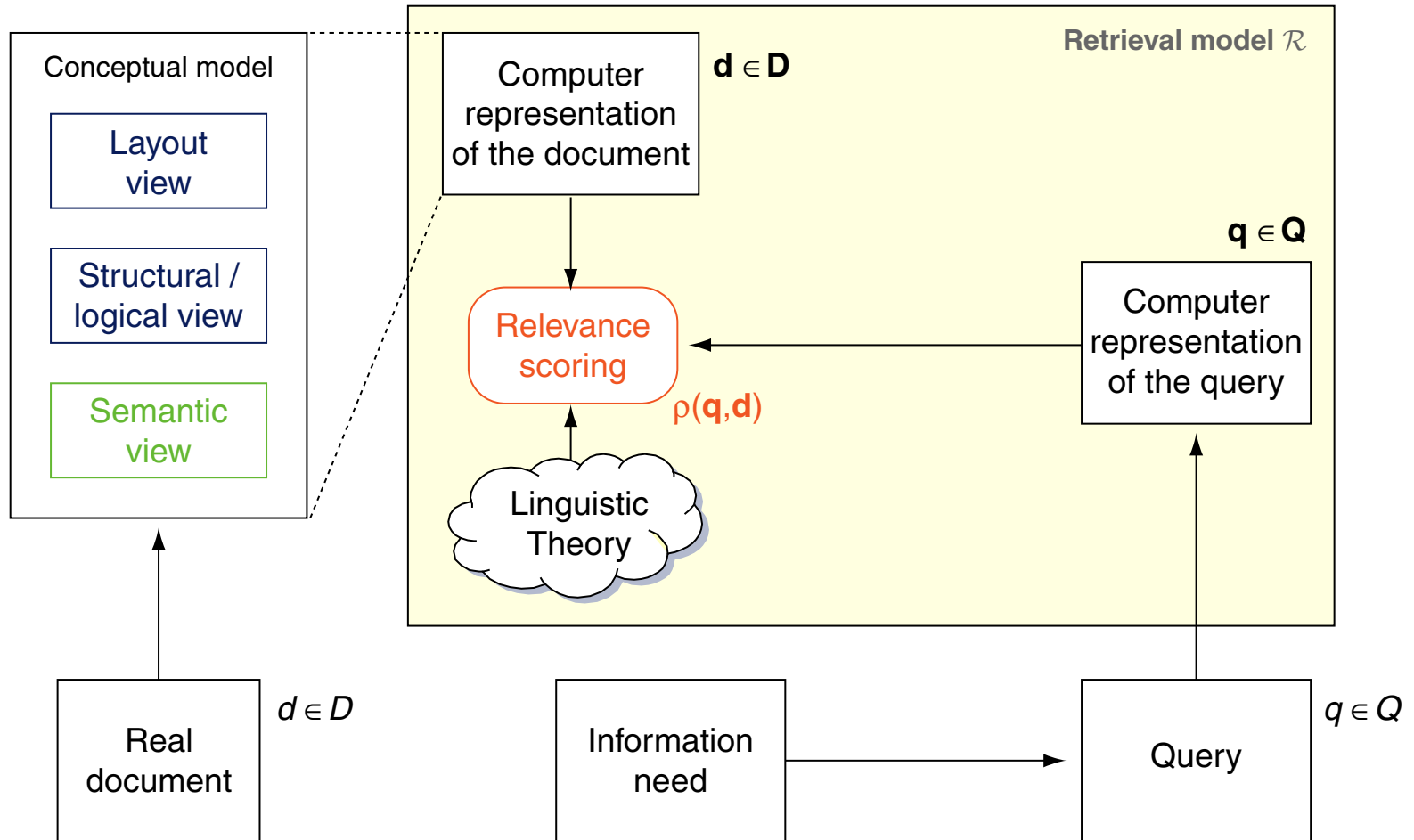
# Overview of Retrieval Models

## Retrieval Models



# Overview of Retrieval Models

## Retrieval Models



# Overview of Retrieval Models

## Definition 1 (Retrieval Model, Relevance Function)

Let  $D$  denote the set of documents and  $Q$  the set of queries. A retrieval model  $\mathcal{R}$  for  $D, Q$  is a tuple  $\langle \mathbf{D}, \mathbf{Q}, \rho \rangle$  defined as follows:

1.  $\mathbf{D}$  is the set of document representations, where  $\mathbf{d} \in \mathbf{D}$  represents  $d \in D$ .  
It may encode information from the layout view, the logical view, and the semantic view.
2.  $\mathbf{Q}$  is the set of query representations.

# Overview of Retrieval Models

## Definition 1 (Retrieval Model, Relevance Function)

Let  $D$  denote the set of documents and  $Q$  the set of queries. A retrieval model  $\mathcal{R}$  for  $D, Q$  is a tuple  $\langle \mathbf{D}, \mathbf{Q}, \rho \rangle$  defined as follows:

1.  $\mathbf{D}$  is the set of document representations, where  $\mathbf{d} \in \mathbf{D}$  represents  $d \in D$ .  
It may encode information from the layout view, the logical view, and the semantic view.
2.  $\mathbf{Q}$  is the set of query representations.
3.  $\rho(\mathbf{q}, \mathbf{d})$  denotes a relevance function, which quantifies the relevance between a query  $q$  and a document  $d$  via their representations  $\mathbf{q} \in \mathbf{Q}$  and  $\mathbf{d} \in \mathbf{D}$ :

$$\rho : \mathbf{Q} \times \mathbf{D} \rightarrow \mathbf{R}$$

The values computed by  $\rho$  are called relevance scores.

$\mathcal{R}$  formalizes a certain **ranking principle**.



## Remarks:

- ❑ A document representation encompasses certain elements and specific aspects of a real document. Examples for document representations include feature vectors, feature trees, and fingerprints.
- ❑ A retrieval model provides the theoretical foundations of how human information needs can be satisfied by drawing information from the three views. Examples for retrieval models include the vector space model, the binary independence model, and latent semantic indexing.
- ❑ An alternative name for a retrieval model is retrieval strategy.
- ❑ Most retrieval models are based on the semantic view of documents.
- ❑ An intensional definition of the sets  $\mathbf{Q}$  and  $\mathbf{D}$  can be given as functions  $\alpha_Q : Q \rightarrow \mathbf{Q}$  and  $\alpha_D : D \rightarrow \mathbf{D}$ . [Fuhr 2004]

# Overview of Retrieval Models

## Ranking Principles in IR

A ranking principle states a criterion and shows that ranking documents by this criterion achieves an objective, usually the maximization of an objective function.

- ❑ Binary user relevance  $rel(d, q) \in \{0, 1\}$  of a document  $d$  to a query  $q$ .  
Objective: Return relevant documents.
- ❑ Semantic similarity  $\varphi(d, q)$  of a document  $d$  to a query  $q$ .  
Objective: Return documents that match the topic of the query.
- ❑ Probability of user relevance  $P(rel(d, q) = 1 \mid d, q)$  of  $d$  to  $q$ .  
Objective: Return documents that satisfy the user.
- ❑ Probability  $P(d \mid q)$  of  $d$  having been generated for the topic of query  $q$ .  
Objective: Return documents that match the topic of the query.
- ❑ Amount of information  $-\log_2 P(d \mid tf(t_1), \dots, tf(t_{|q|}))$  carried by  $q$ 's terms in  $d$ .  
Objective: Return documents that carry much information about  $q$ 's terms.

# Overview of Retrieval Models

## Types of Retrieval Models [\[Amati 2018\]](#)

Logical models: Evaluation for a given document  $d$  and a query  $q$  of the truth of

$$\mathcal{I}(\mathbf{d} \rightarrow \mathbf{q})$$

Algebraic models: Computation of the similarity between  $d$  and  $q$  in a vector space:

$$\varphi(\mathbf{d}, \mathbf{q})$$

Probabilistic models: Estimation of the probability of a user's relevance  $rel$  of  $d$  for  $q$ :

$$P(rel(d, q) = 1 \mid \mathbf{d}, \mathbf{q})$$

Bayesian models: Estimation of the probability of generating  $d$  from  $q$ :

$$P(\mathbf{d} \mid \mathbf{q})$$

Information theoretic models: Computation of the number of bits necessary to code  $tf(t_i)$  many of  $q$ 's  $i$ -th term  $t_i$  in  $\mathbf{d}$  for  $i \in \{1, \dots, |\mathbf{q}|\}$ :

$$-\log_2 P(\mathbf{d} \mid \mathbf{tf}(t_1), \dots, \mathbf{tf}(t_{|\mathbf{q}|}))$$

# Overview of Retrieval Models

## Probability Ranking Principle (PRP) [[Robertson 1977](#), [2009](#)]

If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is **the best** that can be obtained for the data.

# Overview of Retrieval Models

## Probability Ranking Principle (PRP) [[Robertson 1977](#), [2009](#)]

If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is **the best** that can be obtained for the data.

### Assumptions:

- ❑ Relevance  $rel(d, q)$  is a property of a document  $d$  given an information need's query  $q$ , assessable without reference to other documents.
- ❑ Relevance is binary:  $rel(d, q) \in \{0, 1\}$ .
- ❑ Relevance is that of an individual user submitting  $q$ .

Ranking by probability of relevance provably maximizes several objective functions:

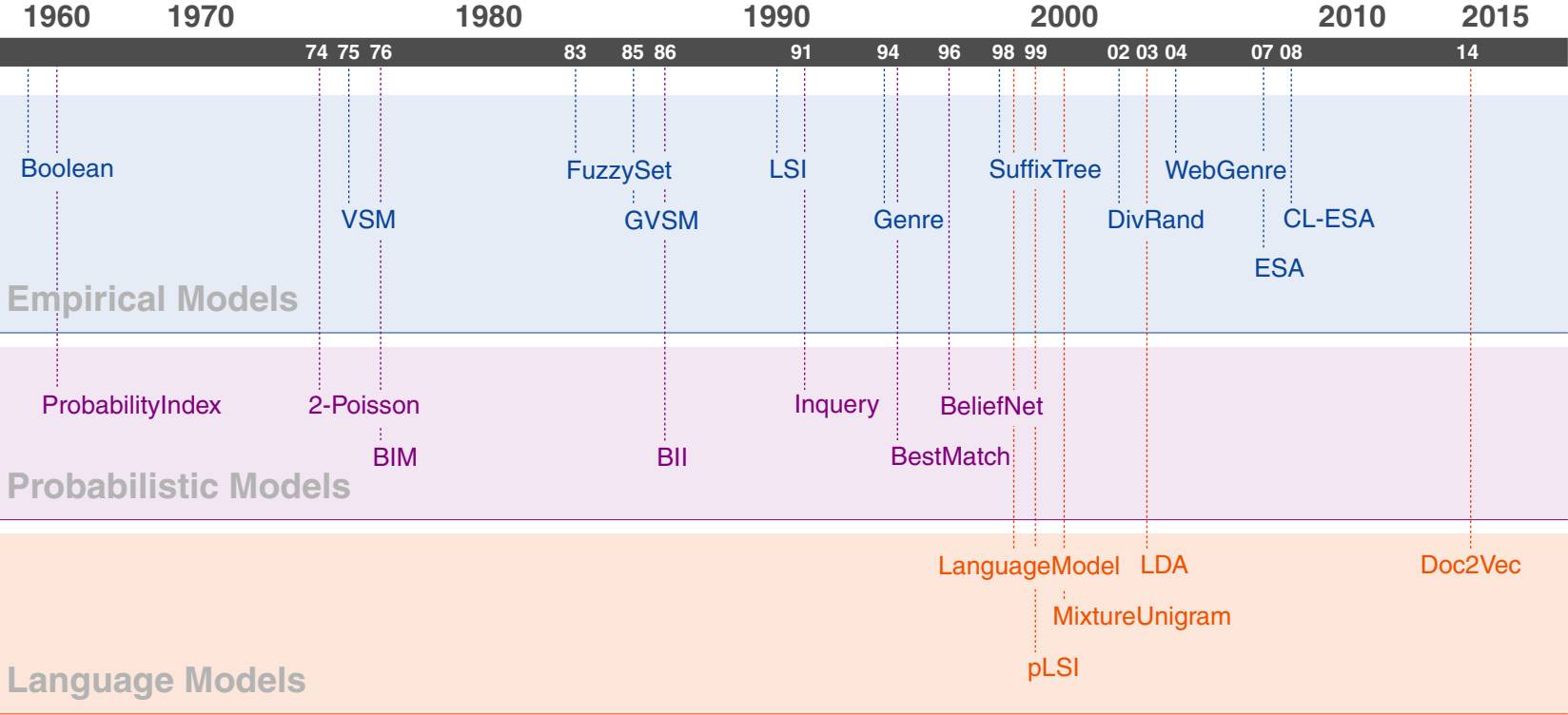
- ❑ Expected recall
- ❑ Expected precision
- ❑ Expected utility

## Remarks:

- ❑ The probability ranking principle has not been shown to hold in general as of yet, but under the above assumptions.
- ❑ William S. Cooper's counterexample considers different users with different information needs who formulate the same query.
- ❑ This can be partially solved by incorporating more knowledge about the user:  
 $P(\textit{rel}(d, q) = 1 \mid \mathbf{d}, \mathbf{q}, \mathbf{u})$ , where  $\mathbf{u}$  is a user model.

# Overview of Retrieval Models

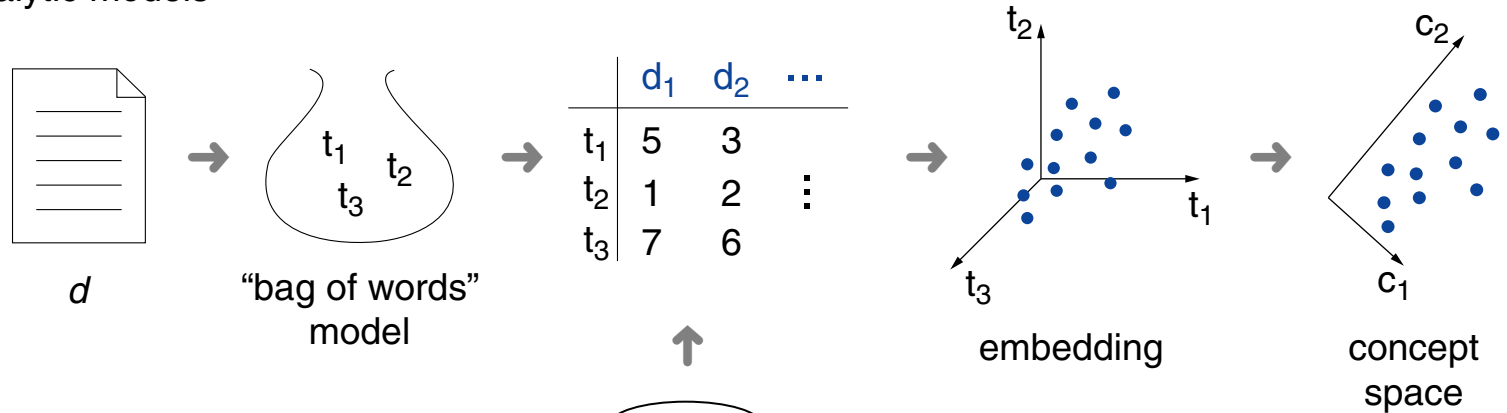
History of Retrieval Models [Stein 2013]



# Overview of Retrieval Models

## Document Modeling

### Analytic models



### Synthetic models / Generative models

