**Course: Advanced Bioinformatics**

**Module title: IR Models**

**Module no. : 168**

For effectively retrieving relevant documents by IR strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporates a specific model for its document representation purposes. The picture on the right illustrates the relationship of some common models. In the picture, the models are categorized according to two dimensions: the mathematical basis and the properties of the model.

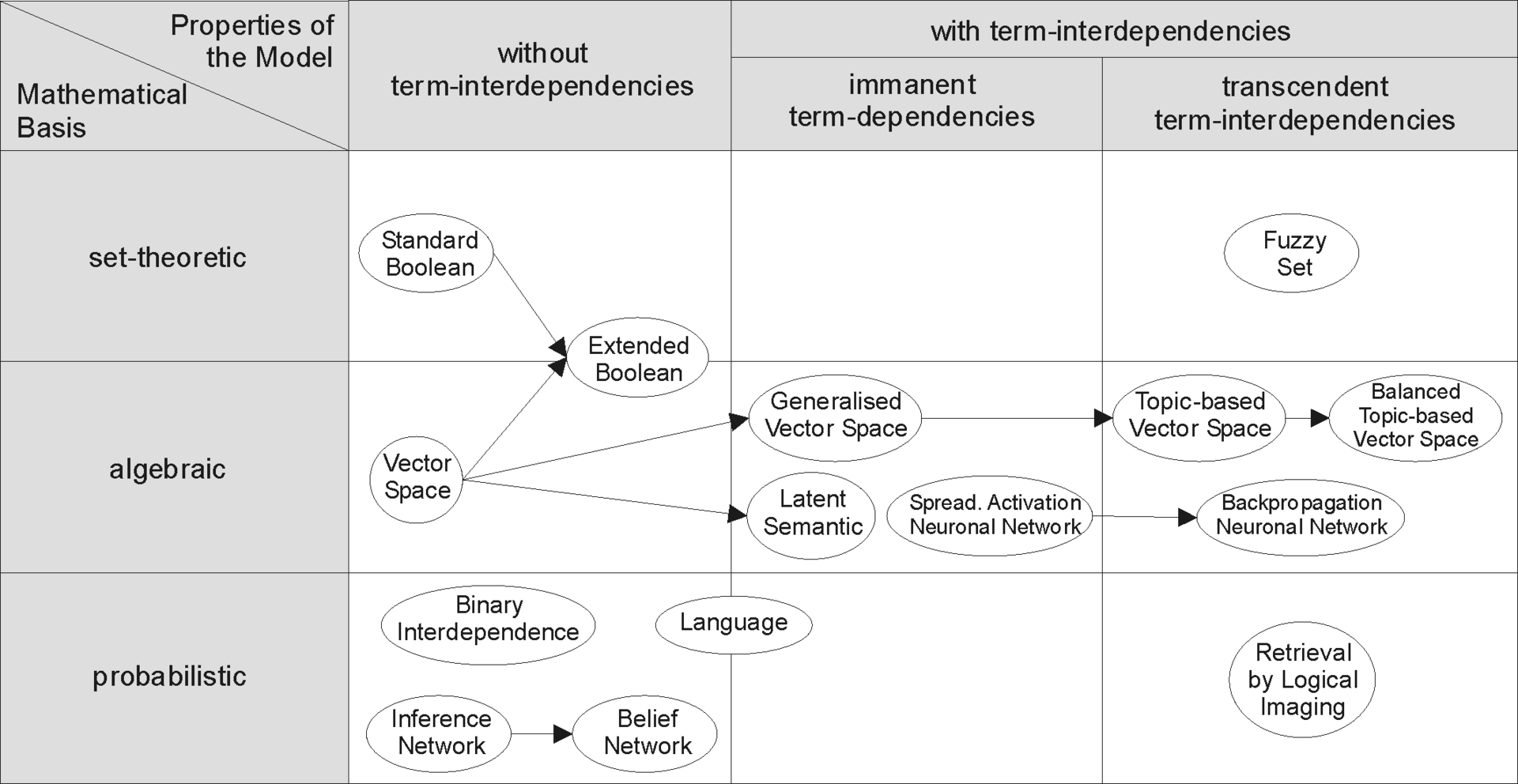
Mathematical Models:

* Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are:
* Standard Boolean model
* Extended Boolean model
* Fuzzy retrieval
* Algebraic models represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.
* Vector space model
* Generalized vector space model
* (Enhanced) Topic-based Vector Space Model
* Extended Boolean model
* Latent semantic indexing a.k.a. latent semantic analysis
* Probabilistic models treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like the Bayes' theorem are often used in these models.
* Binary Independence Model
* Probabilistic relevance model on which is based the okapi (BM25) relevance function
* Uncertain inference
* Language models
* Divergence-from-randomness model
* Latent Dirichlet allocation

Feature-based retrieval models view documents as vectors of values of feature functions (or just features) and seek the best way to combine these features into a single relevance score, typically by learning to rank methods. Feature functions are arbitrary functions of document and query, and as such can easily incorporate almost any other retrieval model as just another feature.

Properties of the model:

* Models without term-interdependencies treat different terms/words as independent. This fact is usually represented in vector space models by the orthogonality assumption of term vectors or in probabilistic models by an independency assumption for term variables.
* Models with immanent term interdependencies allow a representation of interdependencies between terms. However the degree of the interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived (e.g. by dimensional reduction) from the co-occurrence of those terms in the whole set of documents.
* Models with transcendent term interdependencies allow a representation of interdependencies between terms, but they do not allege how the interdependency between two terms is defined. They rely an external source for the degree of interdependency between two terms. (For example, human or sophisticated algorithms.)



Categorization of IR-models