# Why Do We Need Probabilistic Approaches to Ontologies and the Associated Data?

Mehmet Kayaalp, M.D., Ph.D.
National Library of Medicine, NIH, DHHS, Bethesda, MD
Mehmet.Kayaalp@nih.gov

*Conventional ontologies comprise deterministically organized concepts. Certain ontological relations (e.g., those between diseases and causes, symptoms, treatments, or prognosis) cannot be represented faithfully without handling uncertainty. Biomedical ontologies are useful, only if they point to outcomes of the phenomena of interest. Such outcomes are usually associated with probabilistic data. This study is built upon Bayesian probability theory and machine learning where determinism is treated as a special case over a set of probabilistic ontological relations.*

Ontologies are conceptualizations of a particular universe of interest. They map knowledge in a given domain through well-defined constructs. Conventionally, those constructs are based on deterministic logic. Some researchers have also defined a limited set of non-monotonic extensions to deal with uncertainty within the deterministic logic. As of the submission of this study, there is no single Medline® citation with any of the following key phrases *probability theory* or *probabilistic relation* or *statistical relation* along with one of the following key phrases *ontology* or *ontologies* or *knowledge maps*.

Even though we statistically test all experimental information that we gather through biomedical studies, the above observation underlines the fact that the biomedical community has not paid much attention to probabilistic approaches to structuring biomedical knowledge.

Why do we need to adhere to probability theory? Besides a handful of logical constructs such as *is-a*, the rest of the relations necessary to represent biological concepts and phenomena are probabilistic in nature. The characteristics of the *part-of* relation might be sufficient to illustrate the problem.

An organism (biological function) is composed of a number of structural (functional) entities. Such entities are usually related to the organism through the part-of relation. In certain cases, the part-of relation is deterministic in nature. For example, the element carbon must be present in every organic molecule, thus in every living organism. Similarly, every living organism that we know of must have a type of nucleic acid to decode its genetic information. If any of these must-haves is taken out of the composition, the conceptualization of the system would alter drastically.

On the other hand, there are a number of entities that *usually* are part-of an organism, excluding one of which would not alter our overall conceptualization of the organism. For example, soldiers who lose their extremities in the battle are still (classified as) humans. Similarly, a disease (e.g., HIV/AIDS or bird flu) associated with a prevalence may be part-of the characteristics of that population. In short, the part-of relations that are not a subset of must-haves are probabilistic in nature. The probabilistic nature of relations between biologic phenomena (e.g., diseases and their prognoses) is perhaps more obvious.

Why do we need to adhere to the Bayesian approach? The frequentist (as opposed to Bayesian) approach requires a large number of observations before making probabilistic judgments about the outcomes. In most cases, this requirement is impossible to satisfy. The Bayesian approach enables us to quantify beliefs about information through prior probabilities of the interested parties, which may be the designer of an ontology, the user of a model, the researcher of a study, or any combination of them.

The probabilistic approach proposed here has three components: (1) parameters and probabilistic relations, (2) hyperparameters, and (3) data. The first component corresponds to the conventional ontological constructs: entities and their relations. The main difference is that parameters and probabilistic relations are associated with probability distributions. The second component defines those distributions as the beliefs of the interested parties about the underlying distributional characteristics. When those parties are not available, non-informative priors can be used instead. The third component is the data, which can be used to update beliefs and enables its users to make informed decision using their own beliefs about the organization of knowledge and using the associated data.

We can also view component (2) what is believed that exists, (3) what appears (is observed) to exists, and (1) what we can infer that exists (through the combination of 2 and 3). To make this organization operational, we apply machine learning techniques to learn from data and draw inferences.

**Acknowledgement**: The author thanks to Drs. Bodenreider, Humphrey, and Rindflesch for their constructive comments.