# Why Should Machine Learning Require Conceptual Models?

Wolfgang Maass[1], Veda C. Storey[2]

[1] German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus, Germany, wolfgang.maass@dfki.de
[2] Georgia State University, Atlanta, GA, USA, vstorey@gsu.edu

## Abstract

To date, conceptual modeling has been reluctantly used in data science. Therefore, researchers in conceptual modeling should seize the opportunity to explore how data science, generally, and in particular, machine learning, can be improved by using conceptual modeling. This paper outlines the need for conceptual modeling contributions to machine learning and proposes iterative cycles of a conceptualization-data-modeling (CDM) model. Each phase is dominated by processes that focus on knowledge work and technological work. Challenges are identified from both the conceptual modeling and machine learning perspectives.

**Keywords:** data science, machine learning, conceptual modeling, conceptualization-data-modeling (CDM) model

## 1 Introduction

Machine learning algorithms are intended to be general purpose algorithms which means that machine learning models could, in principle, replace traditional software development approaches [1]. For example, it has been shown that an adapted computational neural network (CNN) model can be used for the complex protein folding problem [2]. A key conceptual insight for this work is that the distances between atoms are more informative about the structure than predictions based on direct contact. Local distance information is propagated through the entire protein chain. Inside the architecture, proprietary distance distribution predictions (distogram) are used that have been derived from generic properties of protein chains [2]. From a conceptual modeling perspective, researchers of the AlphaFold project inherently performed many conceptual modeling tasks without making this transparent in their publication. They built an ontological understanding of the domain based on prior research. A topological perspective enabled them to develop a model that predicts local distances between pairs of local protein molecules. Another conceptual decision was that the entire protein chain was considered as a whole and not split into pieces because of computational resource limitations as in prior work. However, little information is provided on the general

conceptual approach, which reduces the potential for understanding, generalizing and transferring this approach to other domains.

Several questions arise. Can this method be generalized and applied to other domains? Is the data specified so that the results can be tested on other datasets? Which ontology is used? Which data quality metrics are applied? What mechanisms are in place for handling missing data and biases? Are the component types of a protein important and, if so, how do we describe their semantics? Which biological knowledge is used during development? Do experts agree on thresholds and similarity measures? Which data constraints and data invariants apply? What is the process for replicating results?

It is typical for machine learning research to be only accessible by well-trained machine learning experts. Conceptual modeling can help bridge the knowledge gaps between experts and non-experts. Specifically, it can help create "verifiable trust" in machine learning systems, instead of "blind trust" that is threatened by instances of model failure.

In this paper, we propose a concise development model that integrates conceptual modeling and machine learning modeling. This model supports understanding of the interaction between knowledge processes and technical processes related to machine learning that has been mostly opaque, even in highly-visible machine learning publications.

## 2 Conceptual Modeling Contributions to Machine Learning

We propose a *conceptualization-data-modeling (CDM) model*, shown in Figure 1, that embeds the technical data science development process into a conceptual process. Developing software changes from implementing functions line-by-line into iterative development steps controlled by objective functions and parameter adjustments. Based on input data, data engineering is performed that is used for model training and optimization [3]. Data requirements and architectures for machine learning models are not randomly selected, but rather, developed in collaboration with experts. Thus, *conceptualizations* proceed *data* collection and subsequently machine learning *modeling*. These three phases are iterated until a model performance is reached that exceeds a given quality thresholds.

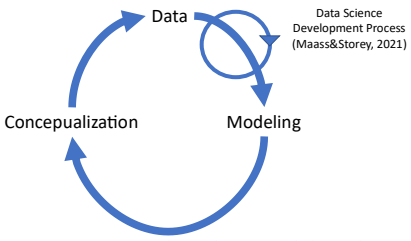


*Figure 1: Conceptualize-data-modeling (CDM) model*

The conceptualization phase integrates the work required for understanding a given domain, investigating a specific problem, and deriving candidate solutions that can be implemented by machine learning. During initial iterations of a machine learning initiative, the development of shared domain knowledge is dominant (*knowledge sharing*) until a level is reached where data scientists have gained sufficient understanding so that the *mathematical and technical understanding* increases in dominance. Both subprocesses (shared understanding and mathematical and technical conceptualizations) develop artefacts for communication within and beyond team boundaries.

Analogous to the conceptualization phase, the data phase consists of two overlapping processes. First, creating and finding data that fulfils data requirements is dominant until it becomes clear which data is actually available with sufficient quality (*data screening*). Many projects fail because of insufficient data access or data availability, data quality, or resource requirements. The second process is *data engineering* that consists of technical data work including data exploration, data preparation, and feature engineering [3]. Data engineering interacts with data screening and becomes more important as soon as the right data of sufficient quality has been collected.

The modeling phase considers the core of a machine learning initiative. However, this phase can only be as good as prior conceptual phases and data phases. The modeling phase consists of two overlapping phases. First, different model types are tested on the given problem and available data (*model exploration*). This can consist of evaluation of a wide spectrum of supervised, semi-supervised, unsupervised, and reinforcement learning approaches. Model exploration is conducted until sufficient knowledge has been gained that separates promising from less promising approaches. The second process is *model optimization* (aka model tuning). This technical phase focusses on one or few promising approaches and tries to find an optimal design for the architecture, hyperparameters, objective functions, heuristics and other parameters. Seminal papers on machine learning focus on results of the modeling phase alone.

## 3 Conclusion

This paper has proposed how researchers in data science can improve their work by incorporating conceptual modeling. A variety of questions and issues were identified that can serve as starting points for a discussion on how data scientists can incorporate conceptual modeling into machine learning.

## References

1. Pérez, J., Barceló, P., Marinkovic, J.: Attention is Turing-Complete. Journal of Machine Learning Research 22, 1-35 (2021)
2. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A.: Improved protein structure prediction using potentials from deep learning. Nature 577, 706-710 (2020)
3. Maass, W., Storey, V.C.: Pairing conceptual modeling with machine learning. Data & Knowledge Engineering 134, 101909 (2021)