Knowledge graph embedding for experimental uncertainty estimation

Edoardo Ramalli and Barbara Pernici
Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

Abstract

Purpose – Experiments are the backbone of the development process of data-driven predictive models for scientific applications. The quality of the experiments directly impacts the model performance. Uncertainty inherently affects experiment measurements and is often missing in the available data sets due to its estimation cost. For similar reasons, experiments are very few compared to other data sources. Discarding experiments based on the missing uncertainty values would preclude the development of predictive models. Data profiling techniques are fundamental to assess data quality, but some data quality dimensions are challenging to evaluate without knowing the uncertainty. In this context, this paper aims to predict the missing uncertainty of the experiments.

Design/methodology/approach – This work presents a methodology to forecast the experiments' missing uncertainty, given a data set and its ontological description. The approach is based on knowledge graph embeddings and leverages the task of link prediction over a knowledge graph representation of the experiments database. The validity of the methodology is first tested in multiple conditions using synthetic data and then applied to a large data set of experiments in the chemical kinetic domain as a case study.

Findings – The analysis results of different test case scenarios suggest that knowledge graph embedding can be used to predict the missing uncertainty of the experiments when there is a hidden relationship between the experiment metadata and the uncertainty values. The link prediction task is also resilient to random noise in the relationship. The knowledge graph embedding outperforms the baseline results if the uncertainty depends upon multiple metadata.

Originality/value – The employment of knowledge graph embedding to predict the missing experimental uncertainty is a novel alternative to the current and more costly techniques in the literature. Such contribution permits a better data quality profiling of scientific repositories and improves the development process of data-driven models based on scientific experiments.

Keywords Uncertainty prediction, Data uncertainty, Data quality, Data quality management, Data uncertainty management, Experimental data, Experimental measurement, Uncertainty prediction

Paper type Research paper

1. Introduction

Experimental data (also experiments in the following) are fundamental to generate chemical–physical predictive models. Such models predict complex systems leveraging chemical–physical equations that describe the domain phenomena. However, some of them are still challenging to explain with theory (i.e. chemical–physical equations), and the experiments, with their observations, can provide phenomenological evidence about a domain setting. This information can be used to refine and validate a model. During model validation, the model's predictions are compared against the experimental data, estimating the predictive model performance. For these reasons, chemical–physical predictive models often are data-driven models (Pelucchi *et al.*, 2019). Experiments, unlikely other types of data such as social media, are rare and expensive in terms of time and cost to collect. An experiment measures physical properties in a given domain setting. They are a particular kind of data because they record physical measurements inherently affected by experimental uncertainty, also known as experimental error (Ramalli *et al.*, 2021b). This is usually obtained by repeating the experiment under the same conditions. Multiple sources of the same fact are hence used to build a *ground truth*, which is compared with the experimental data to compute the difference and thus estimate the uncertainty (Moffat, 1985). Unfortunately, many experiments lack to report experimental uncertainty due to the cost of replicating them (Dai *et al.*, 2019). Similarly, old experiments are more likely to be imprecise, hence with bigger uncertainties, due to the use of old instruments to perform the measurements. However, it is unlikely that the community will invest in

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/2398-6247.htm**

Information Discovery and Delivery
51/4 (2023) 371–383
Emerald Publishing Limited [ISSN 2398-6247]
[DOI 10.1108/IDD-06-2022-0060]

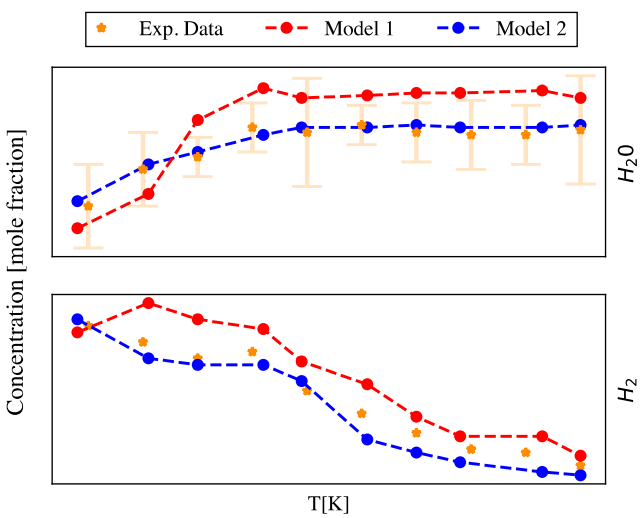
This paper part of special section "Information and data quality for intelligent systems", guest edited by Junhua Ding, Haihua Chen, Lei Li and Ismini Lourentzou.

The work of ER is supported by the interdisciplinarity PhD project of Politecnico di Milano.

Received 30 June 2022
Revised 24 October 2022
21 December 2022
Accepted 21 December 2022

repeating an experiment even if it is not so reliable. Often experiments are one of a kind. In both cases, experiments with or without, with more or less uncertainty, are still valuable sources of information for the predictive model development. Conversely, tasks such as model validation can not be appropriately performed if uncertainty is unavailable. Figure 1 shows an example from chemical engineering where the experimental uncertainty represented by the error bars is a discriminating factor (H_2O plot) to establish if the model predictions are reasonable or not, i.e. the model predictions are inside the experimental error bars. Without them (H_2 plot), it is difficult to guide the model development, thus determining if "model 1" is better than "model 2." Other approaches rely on multiple experiments in similar conditions to estimate the uncertainty (Dai *et al.*, 2019). However, in most of the applications, it is challenging to properly define a similarity between experiments in highly multidimensional and not linear domains. Moreover, these methodologies still rely on multiple experiments to predict the missing uncertainties. Most of the time, the amount of experiments is limited, and most are without uncertainty. This work proposes a new methodology to estimate the missing experimental uncertainty using knowledge graph embedding and the available data. Knowledge graphs, in fact, can represent a data set of experiments given an ontology, and they are easily extensible to include different facts. The proposed methodology leverages three facts: first, predictive models, even if they are affected by epistemic uncertainty (Der Kiureghian and Ditlevsen, 2009), represent more or less faithfully the domain; thus, they can be used to build a ground truth. Second, experiments in similar conditions should report similar values. Learning an embedded representation of the knowledge graph leverages this fact and unties the constraint of defining a distance or similarity between the experiments.

Figure 1 The relationship between experimental data (in orange), simulated data (in blue or red) and experimental uncertainty (error bars). Experimental data without uncertainty (H2 plot) do not allow to establish if the predictions of Model 1 are better than Model 2. Instead, in the case of H2O, it is possible to verify whether model predictions are correct immediately. The data are collected from the scientific repository of SciExpeM (Ramalli *et al.*, 2021b)



Finally, experiments come together with metadata that describes additional details, such as the authorship or the instruments used. It is reasonable to think that sometimes the aleatoric uncertainty (Der Kiureghian and Ditlevsen, 2009) can have a systematic part due to, e.g. a wrong calibration of the instruments of a specific laboratory. In summary, knowledge graph embedding learns hidden, systematic and complex relationships (facts) between the metadata of the experiments and the present uncertainties to predict the missing ones.

Fortunately, in the past decades, there has been an increasing tendency to share data among the scientific community in many scientific sectors. Consequently, many data ecosystems have been created to collect, store and analyze data. These systems are even more critical when dealing with experiments. Data ecosystems play a central role in managing data, de-facto establishing what can be discovered from them (Allan *et al.*, 2012). Data ecosystems offer many functionalities. One of them is related to the data quality evaluation of the repository to provide reliable data to create accurate predictive models. This work briefly discusses which data quality dimensions should be considered and how they should be addressed in the case of a repository for scientific data. In particular, it combines automatic with human-in-the-loop approaches to ensure a certain data quality level. Predicting the missing uncertainty is fundamental to conduct a data profiling of the experimental repository properly. This work validates and studies the new proposed methodology using multiple scenarios using synthetic data. Finally, the procedure is applied to a real-case scenario of chemical kinetics data that are highly affected by missing uncertainties. This proof of concept demonstrates that with enough structured information about the experiments and their uncertainty, it is possible to infer the missing experimental uncertainty. Researchers can find this methodology to predict the missing uncertainties of a scientific repository fundamental. It allows a more effective and straightforward data quality profiling and predictive model development process in many scientific applications.

The structure of this paper is the following. We discuss related work and open problems in Section 2. Section 3 shows how much the experimental uncertainty is fundamental for the data quality profiling of a scientific repository. Then, it presents the proposed methodology. Section 4 shows how the methodology could be effective with a number of scenarios and a real-world case study. Finally, Section 5 summarizes the conclusions and presents future work.

2. Literature review

Uncertainty is a widely used word to describe, in general, a lack of knowledge about the comprehension and description of phenomena (Council, 1990). In the domain of automatically generated models with machine learning techniques, aleatoric and epistemic are two macro typologies of uncertainty (Der Kiureghian and Ditlevsen, 2009). Aleatoric uncertainty is related to the intrinsic randomness of the observed phenomena, e.g. the measurement error in the experiments or fuzziness in an image. Epistemic uncertainty can be associated with the use of a reduced number of variables to represent a complex domain in a predictive model. The predictive model simplifications lead to uncertainty in the predicted values (Der Kiureghian and Ditlevsen, 2009). It is possible to classify uncertainty in more specific categories (Bell, 2001). For

example, the data uncertainty due to the finite precision of instruments leads to random errors or does not account for other factors, such as the instrument drifts that, instead, lead to systematic errors, as well as sampling errors (Aggarwal and Philip, 2008). Uncertainty, when provided, should be adequately processed with uncertain data management methods (Agrawal *et al.*, 2006) and algorithms (Qin *et al.*, 2009; Cormode and McGregor, 2008). In these cases, there are two main ways to represent the data uncertainty (Xu *et al.*, 2014). First, uncertainty can be modeled as a probability distribution of the data rather than deterministic facts. The second is to provide data together with statistical information, such as the average and standard deviation.

Data uncertainty has a central role in the development process of a data-driven predictive model based on experimental data. To mitigate the garbage in, garbage out (GIGO) effects (Kim *et al.*, 2016; Lidwell *et al.*, 2010), experimental uncertainty is fundamental to understand if the data is reliable enough to guide the model development, but often it is not reported (Dai *et al.*, 2019). The lack of discussion about the uncertainty of the data in the experimental sector is mainly due to two factors: either the impossibility of replicating the experiment in the same conditions or the high production cost (Dai *et al.*, 2019).

To guide the model development process through the data, it is necessary to estimate the experimental uncertainty (Hills, 2006). As reported in Peters (2001), there are mainly two methodologies to quantify the uncertainty of the measurements. If it is possible to replicate the measurements inexpensively, the experimental uncertainty can be quantified through experimental campaigns (Moffat, 1985), where a sufficient number of measurements is needed to estimate uncertainty accurately (Hsu *et al.*, 2009). The measurement average is the best estimate for the value to be reported, and the standard deviation is its uncertainty (Peters, 2001). The second methodology instead leverages the Taylor series expansion, and it is mainly used when it is not possible to measure a quantity directly (Wilson and Smith, 2013). When it is not possible to replicate an experiment, a novel approach to estimating data uncertainty leverages the idea that the dependent variable changes smoothly when each independent variable change a little while others are kept constant. This assumption allows representing the relationship between the dependent and independent variables with regression models, and their residuals can be used to estimate the uncertainty of the dependent variable (Dai *et al.*, 2019). The challenges of such an approach are related to the limited availability of data in similar conditions (i.e. independent variables) and the generality of the assumption that even a slight change in the dependent variables corresponds to a little change in the independent one. Finally, a naive approach is to use default domain values (Olm *et al.*, 2014) to complete the missing experimental uncertainty. However, the uncertainty could be much higher than the suggested default value (Ramalli *et al.*, 2021b).

Data ecosystems collect and manage many data types from different sources to produce new knowledge (Cui *et al.*, 2020; Ramalli *et al.*, 2023). In the past years, scientific repositories and data ecosystems have been increasing in different fields (Blaiszik *et al.*, 2019; Nacházel *et al.*, 2021; Ramalli *et al.*, 2021b). Their data management capabilities and services define what can be discovered (Allan *et al.*, 2012). These repositories, if, on the one hand, incentivize the reuse and proliferation of data, on the other hand open new challenges related to data management (Cui *et al.*, 2020), such as data quality (Oliveira and L'oscio, 2018), diversity (Ramalli and Pernici, 2021), integration (Brodie, 2010) and transparency (Geisler *et al.*, 2021). Having all data in a centralized system, if the data are not adequately managed, can lead to a fast propagation of errors within the system and to the data applications (Fan and Geerts, 2012).

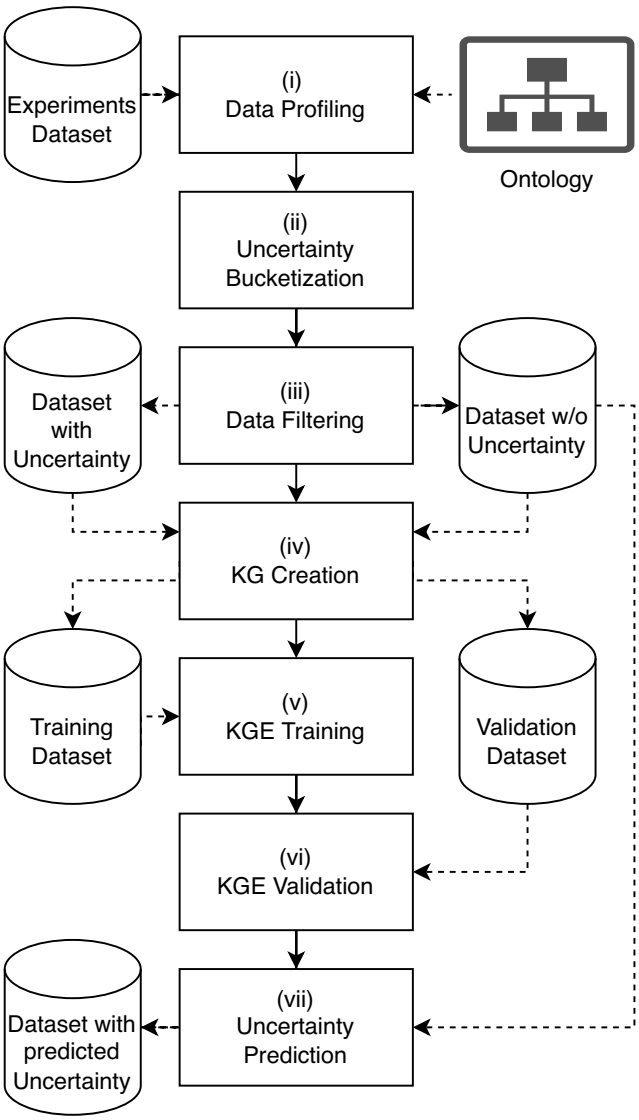
Metadata is known to be used to model the data quality (Rodríguez and Servigne, 2013), and uncertainty can be seen as another metadata of the experiments. This relationship makes it possible to link uncertainty to the data quality (Comber *et al.*, 2006). In the general case, the relationship between data, data quality and uncertainty is linked by data profiling (Naumann, 2014): current data quality reports are imprecise because they lack a complete description of data uncertainty (Comber *et al.*, 2006). Moreover, data quality indicators, independently from the sophistication needed to account for the data uncertainty, do not account for the adequacy of the data for a given goal (Coulon *et al.*, 1997). Metadata can be used to derive an ontology of the data (Chen and Plale, 2012; Schuurman and Leszczynski, 2006), and knowledge graphs (Ehrlinger and Wöß, 2016) can represent ontologies for scientific data (Farazi *et al.*, 2019). The ontology helps in profiling the data and measuring the data quality effectively (Khokhlov and Reznik, 2020).

Knowledge graph embedding is a branch of machine learning that studies how to learn an embedded representation of a knowledge graph. With such representation, this paper shows how it is possible to apply machine learning tasks, such as "link prediction," to infer new knowledge about two entities in the graph (Dai *et al.*, 2020; Wang *et al.*, 2017), focusing on experiments and their uncertainty value.

3. Methodology

This section presents the overall methodology for predicting the missing experimental uncertainties using knowledge graph embedding according to the following steps, also depicted in Figure 2. The methodology is general, and every practitioner who wishes to apply in their domain has to follow the following steps. To start, the researcher needs a scientific repository of experiments with missing uncertainty values and an ontology that describes the experiments. Uncertainty is assumed to be a property of the experiment ontology. The experiments are profiled mainly for two purposes. First, to assess the data quality, thus ensuring the highest data quality possible for the forthcoming knowledge graph generation. Second, to quantify the diversity of the uncertainty values. Generally, the knowledge graph embedding models can not predict continuous values. Therefore, it is necessary to select a limited but representative and exhaustive set of possible uncertainty values (or buckets) according to the application domain and transform the input data set uncertainties using bucketization. With bucketization, similar (close) values are associated with the same bucket. After separating the experiments with uncertainty from the ones without uncertainty, according to the provided ontology, it is created a knowledge graph of the input experimental data set (Ehrlinger and Wöß, 2016), and

Figure 2 Steps of the proposed methodology to predict the missing uncertainties of experimental data



randomly generated a training, validation and test set in the measure of 80, 10 and 10% of the original data set, respectively. In the end, an embedded representation of the knowledge graph is learned and then validated. If the validation results are satisfactory, the embedded model is used to predict the missing uncertainties with the task of link prediction.

Section 2 presents methodologies to predict the experimental uncertainty based on the assumption that multiple experiments exist in similar or in the same conditions to carry out a robust statistical analysis. However, as already said previously, it is rare to have multiple data in similar conditions, and therefore these, methodologies are not applicable.

It is necessary to rely only on the already available information. With the link prediction task, knowledge graph embedding learns and predicts the missing uncertainties based on the hidden evidence in a given data set. If there is a systematic relationship between the experiment metadata, i.e. the experiment ontology properties, and the experimental uncertainty, the embedded representation of the knowledge graph leverages this fact to predict the missing uncertainty values. The validation result of the embedding model over the task of predicting the missing uncertainties confirms if such hidden relation exists. Let us consider, e.g. the case of a systematic instrument drifting in the equipment of a particular laboratory. The equipment type, the laboratory and the uncertainty are properties of the experiment ontology that have been transformed into knowledge graph entities during the knowledge graph creation. Moreover, knowledge graph embedding can be seen as an extension of the previous methodologies: if multiple experiments exist in the same or similar condition, their embedding will be similar, and therefore, the uncertainty predicted by the embedding model will take advantage of this fact. Finally, the ontology and, thus, the knowledge graph, can be easily extended to include additional facts that can be leveraged to predict the missing uncertainty.

The following Section 3.1 describes in more detail the data quality dimensions of interest in a repository for scientific data, which is the procedure to account for them, and how uncertainty is linked to data quality profiling. Section 3.2 presents the ontology of experimental data using a knowledge graph that has been used to validate the proposed methodology. In the end, Section 3.3 introduces how embeddings are computed and used to estimate the missing uncertainties, together with the metrics to assess the embedding quality.

3.1 Data and quality management

The predictive model development process is a complex procedure. It is necessary to analyze many data of different kinds to build it accurately. In particular, there are four types of data: experiments, simulations, models and analysis results. These data, due to their volume and complexity, need to be stored and managed in a data ecosystem that also incentivizes the collaboration and sharing of information among different research institutions.

The data ecosystem manages these four types of data, which are linked together as follows. Models describes real phenomena and are founded on chemical–physical equations, but with the increasing amount of data, the models are more and more data-driven. For this reason, the models are becoming increasingly data-dependent because experimental data are used to measure the goodness of the model prediction. At the same time, using the models, it is possible to generate, neglecting the computational cost, as much simulated data (or simulations) as necessary. These data are model predictions about the behavior of a system in a particular condition. The experimental data are compared with the corresponding simulations to assess the model's prediction quality, generating *performance analysis data*. Moreover, as the model represents reality more or less precisely, they can be used to *validate an experiment*. An experiment that is far away from the model prediction suggests that it is a possible outlier that needs to be investigated. In addition, multiple models developed independently from each other can be used to generate simulated data and estimate the uncertainty of the experimental data, constructing a reasonable ground truth, even if they are affected by (epistemic) uncertainty. Finally, analysis data can be used to *guide the model development improvement or the*

experiment design, i.e. to discover which portions of the domain are not covered by experimental data.

Therefore, as the four types of data are connected with each other, it is more critical that the data stored inside the repository retain certain data quality. Otherwise, spreading wrong information could negatively and rapidly impact the data ecosystem.

This work focuses on the data quality dimensions related to the experimental data, which are the most affected by errors among the four data types. Experimental data are a composition of measurements about a property together with a collection of metadata that provides details about the circumstances of an experiment, such as environmental conditions, instruments, authorship and year of publication.

Experimental data, in terms of data quality, should be checked a priori during the insertion in the data ecosystem, and if they are not compliant with the predefined data quality rules for each dimension, they should not be accepted in the system. In doing so, immediate feedback is provided to the user inserting the data, giving the possibility to correct errors. The procedure to assess the data quality constraints relies on a combination of automatic and human-in-the-loop methodologies.

There are hundreds of data quality dimensions that measure different aspects of the quality of data. When dealing with experimental data, regardless of the application domain, completeness, consistency and accuracy are always of interest according to the fitness for use concept (Wang and Strong, 1996). Unlike many other applications, timeliness as a data quality dimension is often not of interest. It is sporadic that two experiments are performed in the exact experimental condition to update the old value. This is related to the infeasibility of repeating an experiment. First, because it is practically challenging to replicate environmental conditions exactly; second, experiments are expensive, and it is unlikely that other researchers invest in an experiment that has already been investigated.

Completeness. In a domain, it is possible to know which metadata (or ontology properties) that describe the experiment are mandatory and in which conditions. For example, in every experiment is common that the unit of measurement is mandatory. Therefore, it is sufficient to specify a collection of rules that check the completeness of the experiment's metadata.

Consistency. As in the case of completeness, in each domain, there are implicit rules that need to be made explicit and implemented automatically. A typical example is the concordance between the type of the measured property and the unit of measurement. Just subsets of all existing units of measurement are possible for a type of property.

Accuracy. Accuracy is by far the most challenging data quality dimension to evaluate. In the case of experiments, because there is no ground truth and uncertainty is always present in the measurements, it is hard to verify whether the measurements are correct or inaccurate. To overcome this limitation, instead of carrying out a single experiment, a campaign of experiments is performed, where the measurements are repeated several times, but it comes with the infeasibility of replicating experiments. The average of the measurements is the value, and the standard deviation is how much confident we are about that value, or in other words, its uncertainty. The best approach to measure the accuracy of an experiment is to run the corresponding simulation automatically and check if the simulated data are not too far from the experimental ones. In cases in which there is a significant difference, the intervention of a human expert is needed because it is hard to disambiguate automatically whether the data or the model are wrong. Otherwise, checking *a priori* all the experiments manually would be unsustainable for the number of human resources needed.

Uncertainty connects all three data quality dimensions. In the scientific domain, uncertainty is a fundamental element that permits to profile their data quality. For experimental data, uncertainty is another metadata of the experiment; therefore, if available, the repository is" more complete." Instead, if two experiments are in the same conditions (but with different authorship) and without uncertainty, either they report the same observation or both or one of them is wrong. Therefore, uncertainty matters in terms of consistency between experiments and similarly for accuracy. In other words, uncertainty gives a margin of error that allows a more correct assessment of whether an experiment is similar to simulated data with some margin and thus is not just an in-and-out punctual value comparison.

3.2 Knowledge graph

Experiments are a collection of metadata and reported observations about a measured property. An ontology can be used to transform a set of experiments into a knowledge graph. The ontology classes and properties become entities, and the relationships between the classes become the knowledge graph predicates.

Definition 3.1 (knowledge graph). A knowledge graph (KG) is defined as $G = \{E, R, F\}$*, where E, R and F are collections of entities, relations and facts, respectively. A fact is denoted as a triple* $(h, r, t) \in F$ *where h stands for 'Head' (or subject), r stands for 'Relation' (or predicate) and* t *for 'Tail' (or object). In other words, a KG is a list of facts, each representing a truth that connects two entities with a specific relationship. The representation format of the triples is known as resource description framework (RDF)* (Lassila *et al.*, 1999).

It is necessary to carry out several tests to validate the proposed methodology's applicability and understand its strengths and limitations. This work, starting from existing scientific ontologies in the literature (Varga *et al.*, 2015; Farazi *et al.*, 2019), define a metamodel to describe an experiment and hence to create a knowledge graph as shown in Definition 3.1. This ontology accounts for the most popular and general metadata that describe an experiment. The machine learning algorithm will leverage the hidden patterns between these metadata to predict the missing uncertainties, as explained in the following sections. It is plausible that not all the discriminant elements to determine the experiment uncertainty are included in the knowledge graph, but the model embedding validation will notify this limitation. On the other hand, if properties or classes that do not concur with the experiment uncertainty are missing from the ontology, the embedding accuracy will negligibly affect the link prediction of the uncertainty itself (Ramalli *et al.*, 2021a).

In particular, the ontology used in this work (Figure 3) accounts for:

- *Experiment*. It is the ID that identifies an experiment.
- *Author*. It is the first author or the research lab that publishes the experiment. This metadata could be

expanded in other metadata with other bibliography information such as journal, list of authors and affiliations.

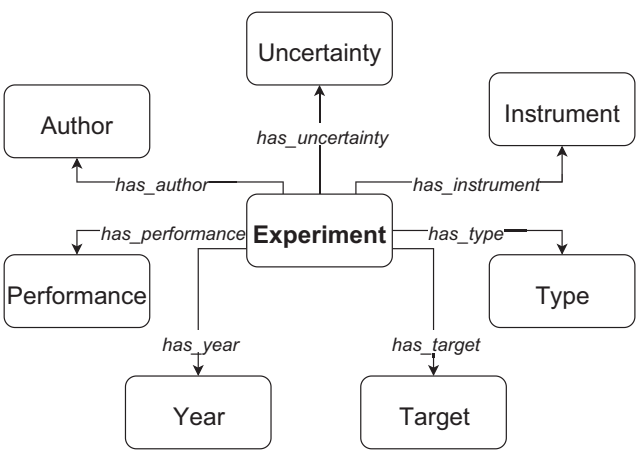
- *Performance*. It represents how much the experimental measurements are distant from the predicted one by a model that is taken as a reference. This performance index is computed by using functional analyses. The values range from 0 to 1, where 1 is the perfect similarity. For this work, the possible values are discretized equally in ten parts from 0 to1.
- *Year*. It is the publication year of the experiment.
- *Target*. It is the object of the experimental investigation. In the general case, multiple properties could be the object of the observation in the same experiment. It is possible to model such a case by adding a new (experiment) entry for each object to the knowledge graph where all the other properties (metadata) are unchanged except for the performance and the uncertainty.
- *Type*. It is the typology of experimental investigation.
- *Instrument*. It represents the instrument used to carry out the experiment. Usually, for each type of experiment, only a subset of the existing instruments is possible.
- *Uncertainty*. It is the (relative) uncertainty of the data, if it is provided. As in the case of the performance index, for this work, the possible uncertainty values are discretized from 0 to 1 with step 0.1. 0 means that it is not possible to determine the experiment uncertainty.

3.3 Knowledge graph embedding

Embedding is the process of representing a complex entity in a lower dimensional space such that entities with similar semantic meanings have close embeddings. Therefore, knowledge graph embedding is the task of creating a knowledge graph representation in a low-dimensional space of size k. Different types of knowledge graph embeddings differ in the representation space, scoring function, encoding models and any other additional information that can be integrated into the embeddings (Rossi *et al.*, 2021; Wang *et al.*, 2021). These characteristics together are known as an embedding model.

The most commonly used family of embedding models uses Euclidean spaces to learn the vector representation of the entities and relationships. To this family of embedding models belongs TransE (Wang *et al.*, 2014), and the more recent RotatE (Sun *et al.*, 2019).

Figure 3 Representation of the metamodel of a typical experiment using a knowledge graph



TransE learns the embedding of the entities and the relationships by interpreting them as a translation on the Euclidian space of dimension equal to the embedding size. Equation (1) reports the scoring function, where $\overline{h}, \overline{r}, \overline{t} \in R^k$ are the embeddings of the head h, relation r and tail t of a triple (h, r, t).

Therefore, the score function of a triple measures the error of the embedding model in representing the mathematical relation between the entities (embedding) and the relationship (embedding) of a triple. In the case of TransE, it measures how much the vector representing the tail $\overline{t}$ of a triple is distant from the vector of the head $\overline{h}$ plus (vector sum) the vector of the relation $\overline{r}$.

Therefore, the purpose of TransE, but in general of every embedding model, is to minimize the loss over a training set of triples T as in equation (2) within a number of given epochs.

Figure 4 shows a visual representation of a possible embedding for two entities and a relation in the case of TransE. Given a triple such as (Milan, is in, Italy), TransE learns the embedding for each element of the triple, such as the loss is minimized:

$$f_{(h,r,t)} = \overline{h} + \overline{r} - \overline{t} \tag{1}$$

$$Loss = \sum_{\forall (h,r,t) \in T} f_{(h,r,t)} \tag{2}$$

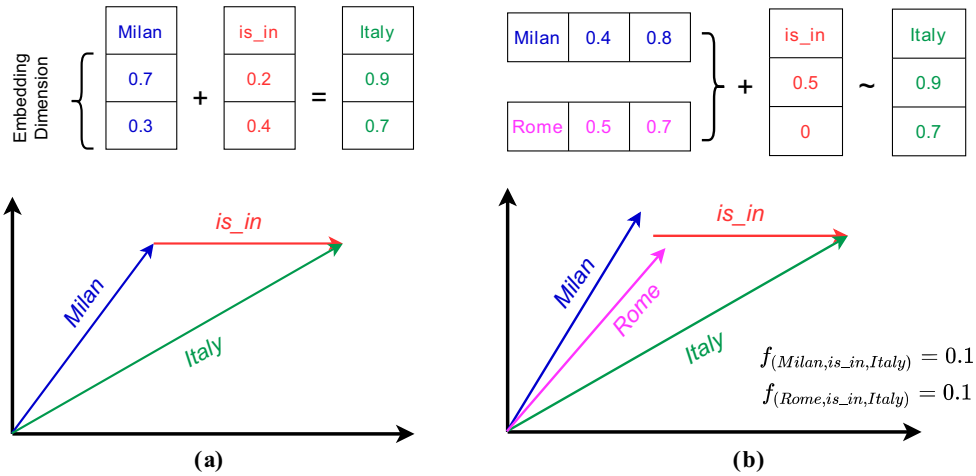
This paper uses RotatE because it is one of the best trade-offs between computational complexity and accuracy of the representation while keeping relational properties, unlikely TransE, such as symmetry (e.g. marriage), inversion (e.g. child and parent) and composition (e.g. my parents' parents are my grandparents). The idea of RotatE is very similar to TransE. It projects the entities and relations into a complex space where the tail entity is reachable from the head entity by a rotation defined by their relationship. In mathematical terms, given a triplet (h, r, t), RotatE learns the embedding $\overline{h}, \overline{r}, \overline{t} \in C^k$ such that it satisfies the mathematical relation $\overline{t} = \overline{h} \circ \overline{r}$, where $\circ$ denotes the Hadamard product.

3.3.1 Link prediction.

With the embedded representation of a knowledge graph, it is possible to use the embedding of the entities and relationships for prediction tasks. The most common are clustering and knowledge completion. The latter is also known as link prediction, i.e. the task of inferring the missing links between entities in the knowledge graph. For the purposes of this work, given the embedding of an experiment and the relationship "has_uncertainty," it is possible to derive which is the most probable embedding associated with an entity representing an uncertainty value, i.e. estimating the uncertainty of an experiment.

Simplifying, the procedure is as follows: given the embedding and the score function of an embedding model, with knowledge completion, it is completed a triple where it is missing one at a time, the head, the relation or the tail. So, given two elements of a triple, it is possible to infer the third one. For example, if the tail in the triple is missing, first, it is necessary to collect the

Figure 4 Two steps of embedding procedure of a KG. Figure 4(a) shows a possible embedding of a triple in the case of TransE. Figure 4(b) depicts how the embedding is computed for multiple triples, minimizing the loss



embedding of the head and the relation. Then, the embedding of all the possible existing (and logically meaningful) tails in the knowledge graph are retrieved. For each possible combination of the head, relation and tail, the link prediction task computes, using the score function of the embedding model, the score and ranks the triples based on the lowest score, i.e. the minimal distance or error. The triple with the lowest score is the best candidate to complete the triple. For this reason, in the proposed procedure, it is necessary to profile the diversity of the uncertainty values. A knowledge graph embedding model can only complete a triple with entities already present during the model's training. Figure 5 shows an example of link prediction using TransE as an embedding model. In this case, we need to complete the triple (*Florence, is_in,?)*, where the tail is missing. After the training of the embedding model, the embedding of the entity *Florence* and the relationship *is_in* are available. Then the link prediction task computes the score of each triple that is generated by substituting the "?" in the (*Florence, is_in,?)* triple, with all the other logically meaningful entities in the knowledge graph that can complete the triple. Each triple is ranked based on the lowest score, and the first ranked suggests the missing element of the triple.

3.3.2 Evaluation metric

Link prediction is used as a benchmark to assess the accuracy of the embedding over a test set of triples Q. In this work, it is used to evaluate the embedding model's predictive capabilities to forecast the missing uncertainties correctly. Therefore, different performance metrics can be defined for this prediction task; one is Hits@N (Definition 3.2).

Definition 3.2 (Hits@N). Hits@N (or H@N) is the proportion of correctly predicted triples within the top N predictions of the embedding model following equation (3):

$$Hits@N = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \begin{cases} 1 & if\ rank_{(h,r,t)_i} \leq N \\ 0 & Otherwise \end{cases} \in [0,1] \qquad (3)$$

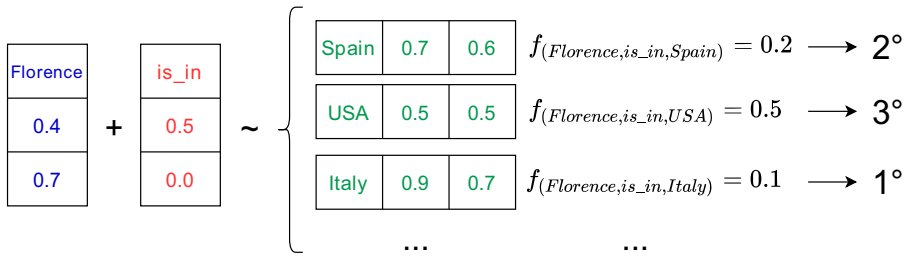
Hits@N has a value between 0 and 1, where higher is better.

Definition 3.3 (Diff). Diff quantifies the average error in the misprediction of the uncertainty. It is the average of difference between the uncertainty value predicted as first top prediction of the embedding model and the real one:

$$Diff = \frac{\sum_{\forall t \in V} |M_1(t) - R(t)|}{k} \qquad (4)$$

More precisely, given a set of triples V ($|V| = k$) where the embedding model mispredicts the first top prediction, *Diff* measures in average, for each triple $f = (h, r, t) \in V$, which is the difference between the value of the first top prediction of the model $M_1(f)$ against the actual value $R(f)$. In such a way, it is possible to understand whether the embedding model is

Figure 5 Example of link prediction of the missing entities in a triple. In this case, given the head Florence and the relationship is_in, the embedding model predict the missing tail with the entity Italy because it has the lowest score, i.e. it has the highest rank



learning the hidden and complex relationship between experiment metadata and uncertainty. For instance, for a given triple f, if the correct uncertainty is $R(f) = 0.5$ for a given experiment, and the first top prediction of the embedding model is $M_1(f) = 0.4$, this is a wrong first top prediction, but the semantic that the embedding model predicts is not so different from the real value. Hence, the embedding model is learning to predict the uncertainty almost correctly.

4. Results and discussion

This section presents the scenarios conducted to study and verify the validity of the new methodology and technologies presented in Section 3. The knowledge graph and the embedding models of the scenarios, together with the source code are available on GitHub [1]. The purposes of the scenarios are the following:

- Set a baseline against which to compare the performance of the other scenarios (Section 4.1).
- (*RQ1*) Determine whether it is possible to learn how to predict missing uncertainty values when it systematically depends on another experiment metadata (Section 4.2).
- (*RQ2*) Evaluate the knowledge graph embedding model's predictive capabilities (of the missing uncertainties) when the depency relationship between uncertainty and experiment metadata is increasingly complex (Section 4.3).

Each scenario uses the ontology in Section 3.2 as a reference for constructing the knowledge graph together with the properties under investigation of the scenario itself. The scenarios are therefore built with synthetic data and are very suitable for performing parametric analyses. In addition, it is possible to easily test the proposed methodology with different sizes of the knowledge graph in terms of the number of experiments, thus of triples paying attention to the increasing training cost.

The scenarios are tested using 1,000 experiments that generate 7,000 random triples but without violating semantic domain constraints between the entities. Table 1 reports in details the number of distinct values for each entity type in the knowledge graph.

Each experiment is randomly (with uniform probability distribution) associated with each type of entity present in the ontology with the proper relationship and feasible entity value. For example, in our application scenario, not all instrument types can be used for every experiment type. It is randomly (due to the computational cost) tested in some scenarios to verify the independence of the methodology from the number of triples if the prediction performances change when are used 50,000 experiments that generate 350,000 triples with a proportional number of entities for each typology. The results suggest that the methodology is independent of the number of triples. The number of possible values for type entities is reported in Section 4.

Table 1 Number of distinct values for each entity type. "Exp." stands for experiment, "inst." for instrument, "uncert." for uncertainty, and "perf." for performance

Exp.	Author	Year	Type	Inst.	Target	Uncert.	Perf.
1,000	50	81	6	5	12	11	11

The training of the knowledge graph embedding model is repeated five times for each scenario. The numerical results, in terms of prediction performance hereafter, are an arithmetic average of five test cases. The list of triples that describe the scenario knowledge graph is randomly divided into three data sets, training, validation and test, respectively, with 80, 10 and 10% of the total triples. The settings for the embedding model are kept constant along all the test cases. In particular, the embedding model is RotatE, with an embedding dimension equal to 64. More details for this setting are in Appendix with Figures A1 and A2. The maximum number of epochs is set to 15,000, with an early stopping on the $H@3$ score over the validation data set computed every 500 epoch with the patience of three steps and delta 5E − 03.

The scenarios are evaluated in two ways: H@N (equation (3)) assesses the prediction capabilities only on the link prediction task for the relationship that connects the experiments to the uncertainty entities. *Diff* (equation (4)) evaluates the average error in the mispredictions.

4.1 Baseline

In the baseline scenario, the knowledge graph is generated following the ontology in Section 3.2. During the generation, consistency rules are kept, such as that experiments belonging to the same author have a plausible publication year from the author's range of activity years. In this case, the uncertainty of the experiment is randomly chosen between 0 and 1. According to this, Figure 6, through the correlation heatmap, shows no correlation between the ontology properties. The embedding result of this configuration of the knowledge graph is in line with expectations. The model predicts the uncertainty of the experiments choosing among 11 possible values. Table 2 illustrates the predictive performance of the embedding model. The predictive performances are not meaningfully above the theoretical limit. For example, $H@5$ indicates the percentage of times the correct uncertainty value to be predicted is in the top five top predictions of the embedding model. Because 11 values are possible, the probability that the correct value is among the top five is about 0.454, which is similar to the value of $H@5$ in the model. Similar considerations about the result of *Diff*. These results support the intuition that if there is no pattern between the uncertainty and the metadata of an experiment, it is not possible to learn how to predict the missing experimental uncertainties. Hence, the uncertainty depends on other, or more complex facts not represented in the knowledge graph. Therefore, even if this methodology depends on the ontology definition of a domain, this preliminary analysis highlights whether all the discriminatory elements to predict the missing uncertainties are included in the ontology, hence in the knowledge graph.

4.2 Research question 1

The baseline scenario demonstrates that if the uncertainty is not linked to any experiment metadata (or ontology property), the best that the embedding model can do is randomly guess the missing uncertainty values. Instead, this scenario verifies if it is possible to learn to predict the uncertainty when there is a systematic relationship between an experiment metadata and the uncertainty. For this purpose, the experiment uncertainty is selected according to the value of another experiment's metadata. More formally, given $X1$ an experiment metadata and $X1$ =

Figure 6 Metadata correlation (Heatmap matrix) in the "Baseline" scenario

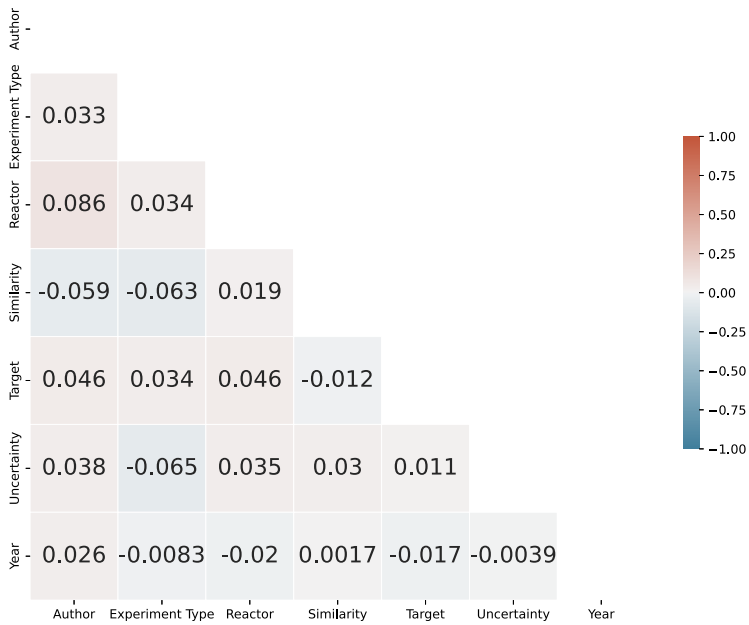


Table 2 Embedding model performance in the "baseline" scenario

Measure	H@5	H@3	H@2	H@1	Diff
Mean	0.465	0.170	0.142	0.087	0.38
Median	0.479	0.167	0.142	0.073	0.38
Max	0.490	0.177	0.142	0.115	0.39
Min	0.427	0.219	0.167	0.073	0.37
SD	0.033	0.021	0.005	0.024	0.01
Var.	0.001	0.001	0.001	0.001	0.00

$\{X1_1,\ldots, X1_n\}$, the possible n value that $X1$ can assume in a domain. Given $U = \{U_1,\ldots, U_k\}$ where $U_j \in [0,1]$, the k uncertainty values present in the knowledge graph. It is need to specify the relationship $\forall X1_i \in X1 \rightarrow U_j \in U$. In this scenario, the relationships are *a priori* randomly chosen from an uniform distribution and kept fixed for the entire knowledge graph generation. Therefore, each possible value of $X1_i$ is always associated with the same uncertainty U_i. Because, in a real case, this association is unlikely to be perfectly strict, it is performed a parametric analysis that increase both the cardinality of U while adding a *run-time* (i.e. during the knowledge graph generation) random (with uniform distribution) but bounded positive or negative deviation (as a random noise) to the relationship between $X1$ and U, such as $X1_i \rightarrow U_j$ becomes $X1_i \rightarrow U_j \pm \sigma$, where $\sigma \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

The parametric analysis results (Figure 7) are shown in Figure 7(a) in terms of $H@5$ score and in Figure 7(b) regarding the *Diff* measure. The knowledge graph embedding model performs well regardless of the number of uncertainty values when the absolute deviation is more strictly bounded. Hence, it is possible to learn a systematic pattern between the experiment metadata and the uncertainty. On the other side, when the absolute deviation is 0.5, it corresponds to the baseline scenario. In fact, because the uncertainty values are bounded $U_j \in [0,1]$, and on average, a deviation of 0.5 from the mean uncertainty value of 0.5 allows all possible values to be associated with the same metadata; hence, there is no relationship between the metadata and the uncertainty.

4.3 Research question 2

The previous scenario shows that knowledge graph embedding can predict the uncertainty of an experiment when there is systematicity between an experiment metadata and the uncertainty itself, even in the face of randomness. This scenario aims to determine whether this methodology can be used when uncertainty depends on an increasing number of metadata. Therefore, how complex the relationship between experiment metadata and uncertainty values can be. A parametric analysis is performed where both the cardinality of U and the number of metadata dependencies increase. In the previous scenario, the relationship between the experiment metadata values and the uncertainties values were expressed with the relation $X1_i \rightarrow U_j$. In this case, instead, the general relation became $(X1_i,\ldots, X4_m) \rightarrow U_j$. Hence, the parametric analysis over the number of dependencies stands for the uncertainty values depending on 0, 1,…,4 experiment metadata value.

The parametric analysis results (Figure 8) are shown in Figure 8(a) in terms of $H@5$ score and in Figure 8(b) regarding the *Diff* measure. As before, when there is no dependency between experiment metadata and uncertainty values, the performance results are similar to the baseline case. When it depends only on one metadata value, it resembles the previous scenario. Instead, if the number of metadata dependencies increase, the embedding performance gets worst but is still quite above the baseline results. This trend is true independently of the number of possible uncertainty values. These results suggest that the knowledge graph embedding model improves the baseline scenario but not as significantly as in other simpler cases.

Figure 7 H@5 (Figure 7(a)) and Diff (Figure 7(b)) result of the parametric analysis over the number of possible uncertainty in the knowledge graph and the magnitude of the random deviation in the relationship between an experiment metadata and its uncertainty

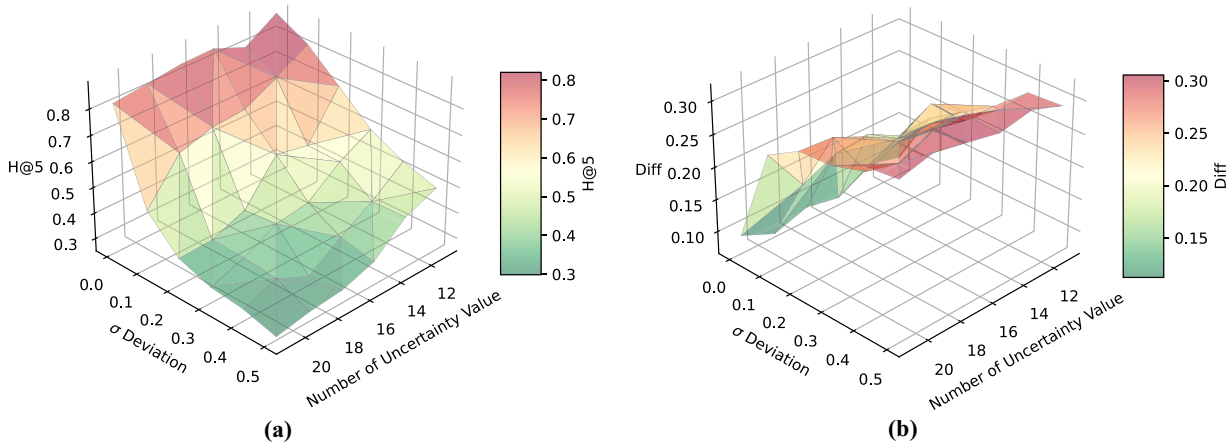
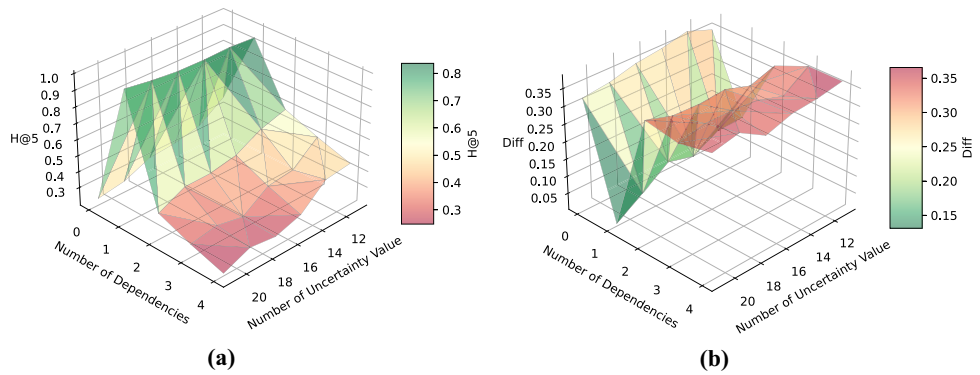


Figure 8 H@5 (Figure 8(a)) and Diff (Figure 8(b)) show the results of the parametric analysis over the number of possible uncertainty in the knowledge graph and the number of dependencies of the experiment metadata and the uncertainty values of an experiment



4.4 Real-world scenario

In this scenario, the same methodology is applied to a data set of chemical kinetics data available in the SciExpeM data ecosystem [2]. A subset of 440 experimental data provided with uncertainty has been collected. The corresponding knowledge graph contains 11,000 triples, where there are six different values of uncertainty. Other information related to the number of possible values for each entity is reported in Table 3.

Also in this case, according to the results in Table 4, the embedding model is able to predict the missing uncertainties even if the task was more manageable. Now, the number of different uncertainties is six instead of 11 as in the general setting of the ontology for these scenarios, which makes unreasonable to measure performance indexes higher than *H@3*. Therefore, the embedding model should guess from a reduced set of possible values. In any case, the results are still exciting and promising.

In the general case, a knowledge graph can be generated in any application in which an ontology can be defined. The more complex the relationship between the knowledge graph entities, the more the learning potentialities of the knowledge graph embedding are leveraged. The above methodology can be used to discover whether there is a dependency between the ontology properties and the uncertainty. In such cases, the embedding model can assess whether it is possible to predict the uncertainty and with which value.

Table 3 Number of different values for each entity. "exp." stands for experiment, "inst." for instrument, "uncert." for uncertainty, and "perf." for performance

Exp.	Author	Year	Type	Inst.	Target	Uncert.	Perf.
440	37	40	5	5	58	6	9

Table 4 Embedding model performance in the "real case study" scenario regarding the estimation of missing uncertainties of the experimental data in the domain of combustion kinetics

Measure	H@3	H@2	H@1	Diff
Mean	0.93	0.88	0.62	0.17
Median	0.93	0.85	0.61	0.16
Max	0.95	0.89	0.62	0.19
Min	0.91	0.87	0.60	0.15
SD	0.02	0.02	0.02	0.01
Var.	0.001	0.001	0.001	0.00

5. Conclusion

Experimental uncertainty is fundamental for the data quality profiling of scientific data, as well as for other predictive model development tasks in which the experiments drive the model development. However, due to the high reporting cost, uncertainty is often missing. In the current state, other methodologies are centered on modeling the available uncertainty or statistically estimating it by relying upon multiple observations in the same domain condition. Because having multiple observations is rare in practice, this work proposes a new methodology to predict the missing uncertainty of experimental data. It leverages the available information and extracts hidden patterns between the experiment metadata and the available uncertainty values. The methodology plans to categorize the existing uncertainties values in n different classes. To predict the missing uncertainties the methodology uses a machine-learning link prediction task. After providing an ontology that describes the experiments, the methodology learns an embedded representation of the knowledge graph that correspond to the provided experimental repository. This methodology is mainly studied with two parametric analyses focused on understanding whether the knowledge graph embedding can learn hidden relationships and how complex they can be to predict the uncertainty values. The results suggest that the embedding model can predict the uncertainty values when there is a relationship between experiment metadata and uncertainty values, even if with random noise. If the relationship is more complex, the embedding model still outperforms the random baseline scenario. The methodology follows the generality principles and can be applied in every scientific domain where it is necessary to predict the missing experimental uncertainty values.

In the future, we plan to study how different predictive models and algorithms perform when compared to knowledge graph embedding. Moreover, we want to investigate how the knowledge graph topology influences the prediction task and, thus, how to leverage better the knowledge graph as it can host multiple ontologies and relationships between entities. We furthermore intend to test a broader set of embedding models to identify which is the most suitable for this specific task. Finally, because knowledge graph embedding has been successfully used for data quality-related activities, we plan to use it for data cleaning and outlier detection tasks.

Notes

1 https://github.com/edoardoramalli/KGE_Exp_Uncertainty

2 https://sciexpem.polimi.it

References

Aggarwal, C.C. and Philip, S.Y. (2008), "A survey of uncertain data algorithms and applications", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21 No. 5, pp. 609-623.

Agrawal, P., Benjelloun, O., Sarma, A.D., Hayworth, C., Nabar, S., Sugihara, T. and Widom, J. (2006), "Trio: a system for data, uncertainty, and lineage", *Proc. of VLDB 2006 (demonstration description)*.

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W.J., Neves, C., Patterson, A., *et al.* (2012), "Omero: flexible, model-driven data management for experimental biology", *Nature Methods*, Vol. 9 No. 3, pp. 245-253.

Bell, S.A. (2001), *A beginner's Guide to Uncertainty of Measurement*, Centre for Basic, Thermal and Length Metrology, National Physical Laboratory.

Blaiszik, B., Ward, L., Schwarting, M., Gaff, J., Chard, R., Pike, D., Chard, K. and Foster, I. (2019), "A data ecosystem to support machine learning in materials science", *MRS Communications*, Vol. 9 No. 4, pp. 1125-1133.

Brodie, M.L. (2010), "Data integration at scale: from relational data integration to information ecosystems", *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, IEEE, pp. 2-3.

Chen, M. and Plale, B. (2012), "From metadata to ontology representation: a case of converting severe weather forecast metadata to an ontology", *Proceedings of the American Society for Information Science and Technology*, Vol. 49 No. 1, pp. 1-4.

Comber, A.J., Fisher, P., Harvey, F., Gahegan, M. and Wadsworth, R. (2006), "Using metadata to link uncertainty and data quality assessments", *Progress in Spatial Data Handling*, Springer, pp. 279-292.

Cormode, G. and McGregor, A. (2008), "Approximation algorithms for clustering uncertain data", Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems, pp. 191-200.

Coulon, R., Camobreco, V., Teulon, H. and Besnainou, J. (1997), "Data quality and uncertainty in LCI", *The International Journal of Life Cycle Assessment*, Vol. 2 No. 3, pp. 178-182.

Council, N.R., *et al.* (1990), *On the Shoulders of Giants: New Approaches to Numeracy*, National Academies Press, Washington, DC.

Cui, Y., Kara, S. and Chan, K.C. (2020), "Manufacturing big data ecosystem: a systematic literature review", *Robotics and Computer-Integrated Manufacturing*, Vol. 62, p. 101861.

Dai, W., Cremaschi, S., Subramani, H.J. and Gao, H. (2019), "Estimation of data uncertainty in the absence of replicate experiments", *Chemical Engineering Research and Design*, Vol. 147, pp. 187-199.

Dai, Y., Wang, S., Xiong, N.N. and Guo, W. (2020), "A survey on knowledge graph embedding: approaches, applications and benchmarks", *Electronics*, Vol. 9 No. 5, p. 750.

Der Kiureghian, A. and Ditlevsen, O. (2009), "Aleatory or epistemic? Does it matter?", *Structural Safety*, Vol. 31 No. 2, pp. 105-112.

Ehrlinger, L. and Wöß, W. (2016), "Towards a definition of knowledge graphs", *SEMANTiCS (Posters, Demos, SuCCESS)*, Vol. 48 Nos 1/4, p. 2.

Fan, W. and Geerts, F. (2012), "Foundations of data quality management", *Synthesis Lectures on Data Management*, Vol. 4 No. 5, pp. 1-217.

Farazi, F., Akroyd, J., Mosbach, S., Buerger, P., Nurkowski, D., Salamanca, M. and Kraft, M. (2019), "OntoKin: an ontology for chemical kinetic reaction mechanisms", *Journal of Chemical Information and Modeling*, Vol. 60 No. 1, pp. 108-120.

Geisler, S., Vidal, M.-E., Cappiello, C., Lóscio, B.F., Gal, A., Jarke, M., Lenzerini, M., Missier, P., Otto, B., Paja, E., *et al.* (2021), "Knowledge-driven data ecosystems toward data transparency", *ACM Journal of Data and Information Quality (JDIQ)*, Vol. 14 No. 1, pp. 1-12.

Hills, R.G. (2006), "Model validation: model parameter and measurement uncertainty", *Journal of Heat Transfer*, Vol. 128 No. 4, pp. 339-351.

Hsu, S.-H., Stamatis, S.D., Caruthers, J.M., Delgass, W.N., Venkatasubramanian, V., Blau, G.E., Lasinski, M. and Orcun, S. (2009), "Bayesian framework for building kinetic models of catalytic systems", *Industrial & Engineering Chemistry Research*, Vol. 48 No. 10, pp. 4768-4790.

Khokhlov, I. and Reznik, L. (2020), "Knowledge graph in data quality evaluation for IoT applications", *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, IEEE, pp. 1-6.

Kim, Y., Huang, J., Emery, S., *et al.* (2016), "Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection", *Journal of Medical Internet Research*, Vol. 18 No. 2, p. e4738.

Lassila, O. and Swick, R.R., *et al.* (1999), "Resource description framework (RDF) model and syntax specification", W3C Recommendation, REC-rdfsyntax-19990222, W3C.

Lidwell, W., Holden, K. and Butler, J. (2010), *Universal principles of Design, Revised and Updated: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach through Design*, Rockport Pub.

Moffat, R.J. (1985), "Using uncertainty analysis in the planning of an experiment", *Journal of Fluids Engineering*, Vol. 107 No. 2.

Nacházel, T., Babic, F., Baiguera, M., Cech, P., Husáková, M., Mikulecky, P., Mls, K., Ponce, D., Salmanidou, D., Stekerová, K., *et al.* (2021), "Tsunami-related data: a review of available repositories used in scientific literature", *Water*, Vol. 13 No. 16, p. 2177.

Naumann, F. (2014), "Data profiling revisited", *ACM SIGMOD Record*, Vol. 42 No. 4, pp. 40-49.

Oliveira, M.I.S. and L¨oscio, B.F. (2018), "What is a data ecosystem?", *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 1-9.

Olm, C., Zsély, I.G., Pálvölgyi, R., Varga, T., Nagy, T., Curran, H.J. and Turányi, T. (2014), "Comparison of the performance of several recent hydrogen combustion mechanisms", *Combustion and Flame*, Vol. 161 No. 9, pp. 2219. 2234.

Pelucchi, M., Stagni, A. and Faravelli, T. (2019), "Addressing the complexity of combustion kinetics: data management and automatic model validation", *Computer Aided Chemical Engineering*, Vol. 45, pp. 763-798.

Peters, C.A. (2001), "Statistics for analysis of experimental data", in Powers, S.E. (Ed.), *Environmental Engineering Processes Laboratory Manual*, AEESP, Champaign, IL, pp. 1-25.

Qin, B., Xia, Y., Prabhakar, S. and Tu, Y. (2009), "A rule-based classification algorithm for uncertain data", *2009 IEEE 25th International Conference on Data Engineering*, IEEE, pp. 1633-1640.

Ramalli, E., Dinelli, T., Nobili, A., Stagni, A., Pernici, B. and Faravelli, T. (2023), "Automatic validation and analysis of predictive models by means of big data and data science", *Chemical Engineering Journal*, Vol. 454, p. 140149.

Ramalli, E. and Pernici, B. (2021), "Know your experiments: interpreting categories of experimental data and their coverage", *SeaData Workshop at VLDB 2021*, 27-33. CEUR Workshop Proceedings.

Ramalli, E., Scalia, G., Pernici, B., Stagni, A., Cuoci, A. and Faravelli, T. (2021b), "Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering", *Frontiers in Big Data*, Article number 663410, Vol. 4, p. 67.

Ramalli, E., Parravicini, A., Di Donato, G.W., Salaris, M., Hudelot, C. and Santambrogio, M.D. (2021a), "Demystifying drug repurposing domain comprehension with knowledge graph embedding", *2021 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, pp. 1-5.

Rodríguez, C.C.G. and Servigne, S. (2013), "Managing sensor data uncertainty: a data quality approach", *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, Vol. 4 No. 1, pp. 35-54.

Rossi, A., Barbosa, D., Firmani, D., Matinata, A. and Merialdo, P. (2021), "Knowledge graph embedding for link prediction: a comparative analysis", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 15 No. 2, pp. 1-49.

Schuurman, N. and Leszczynski, A. (2006), "Ontology-based metadata", *Transactions in GIS*, Vol. 10 No. 5, pp. 709-726.

Sun, Z., Deng, Z.-H., Nie, J.-Y. and Tang, J. (2019), "RotatE: knowledge graph embedding by relational rotation in complex space", *Proc. ICLR 2019*.

Varga, T., Turányi, T., Czinki, E., Furtenbacher, T. and Császár, A. (2015), "ReSpecTh: a joint reaction kinetics, spectroscopy, and thermochemistry information system", *Proceedings of the 7th European Combustion Meeting*, Citeseer, Vol. 30, pp. 1-5.

Wang, Q., Mao, Z., Wang, B. and Guo, L. (2017), "Knowledge graph embedding: a survey of approaches and applications", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29 No. 12, pp. 2724-2743.

Wang, M., Qiu, L. and Wang, X. (2021), "A survey on knowledge graph embeddings for link prediction", *Symmetry*, Vol. 13 No. 3, p. 485.

Wang, R.Y. and Strong, D.M. (1996), "Beyond accuracy: what data quality means to data consumers", *Journal of Management Information Systems*, Vol. 12 No. 4, pp. 5-33.

Wang, Z., Zhang, J., Feng, J. and Chen, Z. (2014), "Knowledge graph embedding by translating on hyperplanes", in Brodley, C.E. and Stone, P., (Eds), *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014*, Québec City, Québec, Canada, AAAI Press, pp. 1112-1119.

Wilson, B.M. and Smith, B.L. (2013), "Taylor-series and monte-carlo method uncertainty estimation of the width of a probability distribution based on varying bias and random error", *Measurement Science and Technology*, Vol. 24 No. 3, p. 35301.

Xu, Y., Fang, X., Li, X., Yang, J., You, J., Liu, H. and Teng, S. (2014), "Data uncertainty in face recognition", *IEEE Transactions on Cybernetics*, Vol. 44 No. 10, pp. 1950-1961.

Appendix. Model parameters

Figure A1 Training loss for each embedding dimension

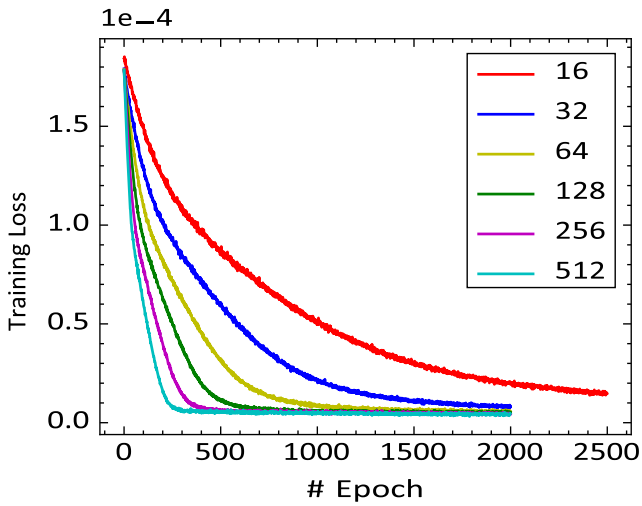
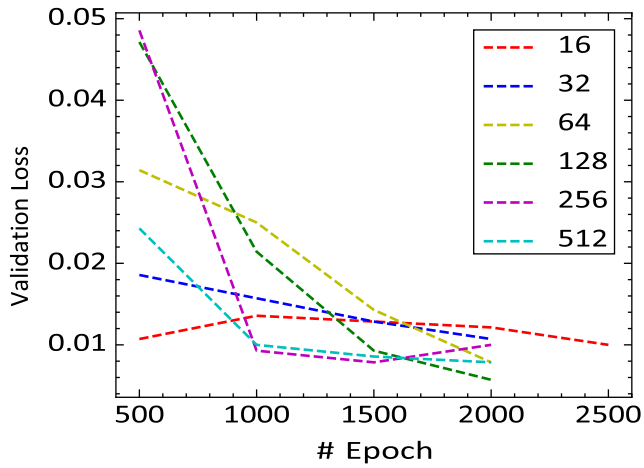


Figure A2 Validation loss for each embedding dimension



Corresponding author

Edoardo Ramalli can be contacted at: edoardo.ramalli@polimi.it