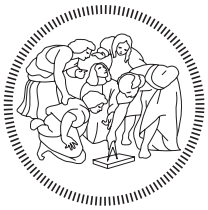


POLITECNICO
MILANO 1863

Department of Electronics, Information and Bioengineering
Doctoral Programme In Information Technology

DATA ECOSYSTEMS AND DATA SCIENCE FOR SCIENTIFIC DATA

Doctoral Dissertation of:
Edoardo Ramalli

Advisor:
Prof. Barbara Pernici

Co-Advisor:
Prof. Tiziano Faravelli

Tutor:
Prof. Davide Martinenghi

The Chair of the Doctoral Program:
Prof. Luigi Piroddi

2023 – XXXVI Cycle

Abstract

Predictive models have a pervasive role in many daily applications. The increasing amount of generated and shared data has recently boosted their development, shifting the model generation and improvement focus towards a data-centric approach. As a result, an information system that manages these data defines what can effectively discovered from them. Predictive models are also used in scientific domains to simulate complex real-world systems, replacing costly and time-consuming experiments. However, the unique characteristics of the scientific data and domain requirements, such as experimental uncertainty, low data quality, and confidentiality, make applying traditional methodologies to share and leverage the data challenging. This interdisciplinary research investigates, as a whole, the development process of a scientific predictive model and how it can be improved by adopting data ecosystem and data science technologies. This thesis focuses on the following requirements: 1) identification of the predictive model development process, classification of scientific data, and their properties, 2) the design of a sustainable data ecosystem to support a quality process, 3) the definition of an effective model evaluation methodology, 4) the use of appropriate data science techniques to guide the improvement and development of scientific predictive models. These requirements and challenges are valid across multiple scientific domains, but the interdisciplinarity of this thesis focuses on a case study of the chemical kinetics field. First, I investigate the current model development process, analyzing the typical steps, the data, and the roles involved. Then, I propose a data ecosystem that offers the necessary services and addresses the unique scientific data proper-

ties and domain requirements as data governance and management aspects while fulfilling the open data guidelines. Finally, the proposed solution is generalized with a set of challenges for designing and adopting sustainable data ecosystems and managing quality data in scientific domains. This thesis presents a systematic, objective, and automatic evaluation methodology for scientific predictive models while handling uncertainties, allowing replicability and awareness of the results with provenance information and fair validation. Finally, it discusses how the results of the model evaluation analysis can inform model improvement and generation. To this end, appropriate data science techniques are used and developed.

Contents

CHAPTER 1

Introduction

1.1 Context

The quantity of generated and shared data is larger than ever [1], and, as a result, many data-driven applications have been developed and used in many aspects of our lives [2], spanning from finance [3] to healthcare [4,5]. One of the primary applications is to use data to make predictive models [6]. Since the application of predictive models is pervasive [7, 8], and the amount of shared data is constantly increasing [9], the generation of a predictive model is shifting from a model-centric approach to a data-centric approach [10, 11]. One of the main challenges in this context is to collect, organize, and effectively extract value from large quantities of data when, for instance, their source, representation, and quality levels are heterogeneous [12, 13], guaranteeing quality data and a sustainable data life cycle [14]. Scientific predictive models (in short, predictive models, or models, in the remainder of this work) are developed on scientific data representing real-world chemical-physical phenomena. Some examples regard meteorology [15], biology systems [16] or chemical kinetics [17]. The latter field, particularly combustion chemistry, is fundamental to the current energy transition agenda since it studies optimizing fuel efficiency and con-

sumption and developing new sustainable and green fuels [18]. To match the ambitious goal of a carbon-free planet [19], it is necessary to speed up and refine the development process of new scientific predictive models and improve their accuracy with ad-hoc data management, data science, and data-sharing techniques [20].

1.2 Motivations

It is, therefore, necessary to improve the development process of scientific predictive models to meet the goal of a sustainable process, more accurate predictions, and faster delivery of predictive models for the current energy transition. At the same time, the promising results of data science and big data management in countless fields can match these needs [21–23]. However, applying such computer science disciplines into the scientific domains, even though particularly promising, is challenging and requires a transformation and adaptation of both sides [20], as explained in the following with more details. In summary, on the chemical kinetics side, the current and consolidated process needs to be rethought. Instead, on the computer science side, a plug-and-play solution using the already available data science and data management technologies is not possible. Therefore, this thesis aims to bridge this gap, explaining the challenges and proposing a solution.

The challenges include:

- Unique characteristics of scientific data. These properties include relatively low volumes, not negligible uncertainty, and heterogeneity in terms of representation formats, sources, and quality levels [24]. Even though a study by the European Commission has predicted that the amount of data generated is increasing from 33 zettabytes in 2018 to 175 zettabytes in 2025, 80% of the data still remains unused[1]. Scientific fields are having a harder time generating and sharing the same amount of data with respect to others, such as social media [25]. Scientific data are much more costly and time-consuming to generate [26], and since they represent real-world phenomena, an infinite number of domain configurations are possible. As a result, scientific repositories are highly sparse and unbalanced. Real-world data are inherently affected by uncertainty. Uncertainty of scientific data depends on a multitude of factors, and assessing the ground truth is not possible [27]. Moreover, ignoring or assuming a constant uncer-

[1] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113

tainty value leads to erroneous conclusions and low-quality data products [28]. Finally, due to their rarity and importance, scientific data are being collected over decades, and all the data are a precious source of information. On the other hand, it is necessary to deal with representation formats, sources, ontologies, and data quality levels that have evolved over the years [29].

- Domain requirements and sustainability. Scientific domains, particularly chemical kinetics, are competitive research fields that have a high impact on the industry [30]. Therefore, even if data sharing is crucial for all data-driven applications, data confidentiality and intellectual property recognition are important requirements for stakeholders to keep the research advantage [31]. Data-sharing platforms have demonstrated successful results, but it is not unlikely that their lifetime is limited due to a business model and system design that is not sustainable [32].

- Engagement. The current scientific model development process is heavily human-based [33]. Therefore, it is subjective, slow, and error-prone. Data management systems can automate part of this process [34], and data science techniques can generate new knowledge [24]. However, since the process is consolidated over decades, gaining trust and users in a new technology is not effortless.

These challenges are not independent factors. For example, to increase the volume of data, many users need to participate in a data-sharing platform, but the low quality of the data can affect their engagement [35]. Furthermore, these challenges are not exclusively limited to scientific data [13]; nevertheless, they take on a distinct significance and possess unique characteristics in the context of scientific data and its applications. Given the abovementioned factors, exploring methods customized explicitly for this domain type becomes pertinent.

1.3 Research Questions

This interdisciplinary research investigates, as a whole, the development process of scientific predictive models and how they can be improved by adopting data management and data science solutions. All the following requirements, derived from the previously discussed challenges (Section 1.2), are investigated in chemical engineering, with chemical kinetics as a case study.

- The *identification of the model development process, scientific data, and their properties* for predictive scientific models. The associate research question is: *"RQ1: Which is the approach to identify all the aspects of a given process, such as the requirements, challenges, roles, services, data, and their properties to be integrated into an automated and systematic information system to improve it?"*.

- The *design of an appropriate sustainable Data Ecosystem (DE) to support such process*. In this respect, the research question is: *"RQ2: Which design choices, services, and functionalities have to be included in a scientific data ecosystem to support the scientific predictive model development process accounting for all the sustainable challenges, domain requirements of the stakeholders, scientific data properties and quality results?"*

- The *definition of an effective model evaluation methodology*. The related research question is: *"RQ3: How can a predictive model be evaluated automatically and systematically in an effective, fair, standardized, and replicable way?"*

- The *use of appropriate data science techniques to guide the improvement and development of a scientific predictive model*. In this respect, the research question is: *"RQ4: How can the result of predictive model validation and the collected scientific data be used to comprehend the model behavior, and how can data science use this information to improve it?"*

The next section details the original contributions to each of these requirements and corresponding research questions. These requirements follow the generality principles and can meet the needs of many scientific fields because they directly stem from the discussed data-related challenges. Each requirement deals with a different perspective and collection of tasks of the predictive model development process. This work focuses on a specific applicative domain and does not propose abstract approaches. Rather, it suggests general methodologies and algorithms that can be evaluated on any dataset in other domains. However, in some cases, the implementation of the methodology, such as the development of a data-sharing platform, is domain-specific, but the investigation procedure and the problem addressed are generalizable and extendible to other domains.

1.4 Original Contributions

The content and some parts of this thesis result from the following publications, where I am either first author, first co-author, or co-author. The publications are published in journals, proceedings of international conferences, or as a book chapter. Some work is currently still under consideration for publication or preparation.

- *Ramalli E, Scalia G, Pernici B, Stagni A, Cuoci A, Faravelli T. Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering. Frontiers in Big Data. 2021 [36].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Pernici B. Challenges of a Data Ecosystem for Scientific Data. Data and Knowledge Engineering. Accepted [37].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Pernici B. Sustainability and Governance of Data Ecosystems. In Proc. 2023 IEEE ICWS. 2023 [38].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Pernici B. From a prototype to a data ecosystem for experimental data and predictive models. In Proc. of the First International Workshop on Data Ecosystems (DEco'22). 2022 [39].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Pernici B. Know your experiments: interpreting categories of experimental data and their coverage. Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores. 2021 [40].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Pernici B. Knowledge graph embedding for experimental uncertainty estimation. Information Discovery and Delivery. 2023 [41].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Bono C, Mülâyim MO, Cappiello C, Carman MJ, Cerquides J, Fernandez-Marquez JL, Mondardini MR, Ramalli E, Pernici B. A Citizen Science Approach for Analyzing Social Media With Crowdsourcing. IEEE Access. 2023 [42].* Implementation.

- *Bono CA, Cappiello C, Pernici B, Ramalli E, Vitali M. Pipeline Design for Data Preparation for Social Media Analysis. ACM Journal of Data and Information Quality. 2022 [43].* Ideation, solution design, implementation, data preparation, and paper writing.
- *Ramalli E, Parravicini A, Di Donato GW, Salaris M, Hudelot C, Santambrogio MD. Demystifying drug repurposing domain comprehension with knowledge graph embedding. In 2021 IEEE Biomedical Circuits and Systems Conference (BioCAS). 2021 [44].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Dinelli T, Nobili A, Stagni A, Pernici B, Faravelli T. Automatic validation and analysis of predictive models by means of big data and data science. Chemical Engineering Journal. 2023 [45].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E. Data Quality, Data Diversity and Data Provenance: An Ethical Perspective. Book chapter in "Improving Technology through Ethics". SpringerBriefs in Applied Sciences and Technology. 2024 [46].*
- *Dilettis M, Ramalli E. Multiple Adaptive Delaunay Optimization. To be submitted [47].* Ideation, solution design, implementation, data analysis, data preparation, and paper writing.
- *Ramalli E, Pernici B, Faravelli T, Deng S. Chemical reaction neural network with element conservation for hydrogen model. Under preparation [48].*

The structure of this thesis and a summary of the main contributions is as follows. Related work is presented in Chapter 2, while a brief introduction to the running scenario of this thesis is described in Chapter 3.

Chapter 4, *"Data Ecosystems for Scientific Data"*, is based on [36–38]. Data management platforms are fundamental to managing a large amount of information [49] and they define what can be discovered from the data [50]. Moreover, even if the available size of data is limited, organizing the data in the same platform enhances the reuse and the value of the data itself [51]. If multiple typologies of data are gathered in the same repository, different kinds of cross-analyses can be applied to extract new knowledge. However, regardless of the application domain, it is necessary to identify the requirements of the stakeholders and the goals of such a platform. In the

case of scientific predictive model development, it is necessary to identify the user roles, the data, and their properties. It is also fundamental to understand how they interact in the process, the challenges posed by integrating scientific data in such a system, which functionalities need to be offered by the platform, and which are missing in the current model development process.

In this respect, the main contributions of this thesis are:

- Identification of the scientific data and their properties in the context of the development of a scientific predictive model.
- Development of the trust-user-data framework to analyze the challenges of developing a data-sharing platform for scientific data.
- Definition of the challenges that need to be addressed for adopting a data-sharing platform to improve the development process of scientific predictive models.

Chapter 5, *"Data Ecosystem architectures for scientific data"*, is based on [36–40]. Once the challenges are defined, this thesis dives into the proposed approach to conciliate data management and data science within the predictive model development process.

For this aspect, the main contributions can be summarized as follows:

- Definition of the DE architecture and the data management and governance aspects that are important to account for to improve the engagement and sustainability of the platform.
- Formalization of the new chemical kinetics development process, the services, the user roles, and how to automatically interpret the semantics of the data in a complex scientific domain.
- Starting from a previous prototype [52], the development of a new DE that is currently used by different research groups in their daily work.

Based on the content published in [36, 37, 40–44], Chapter 6, *"Data Preparation"*, presents the contributions in terms of data preparation. This is an essential step in many applications, particularly the data-driven ones. The contribution of this thesis regarding this aspect concerns the investigation of which data preparation procedures are necessary for proper model validation and improvement, sharing of scientific data, and engagement in the data-sharing platform.

In particular, the following topics were investigated in the thesis:

- How data transparency, in particular, data provenance, can be used to analyze a given workflow to identify which pipeline stage needs improvement, and how the definition of a data provenance model for the predictive model development process can improve the engagement and trust of the users in the data sharing platform.
- The development of knowledge graph embedding to predict the missing uncertainty in the experimental data to validate and improve the scientific predictive models properly.
- The definition of a measure to assess the data diversity of datasets regardless of the domain, in a fast and comprehensive way.

The contributions presented in Chapter 7, *"Model Evaluation and Improvement"*, are the results of the following publications [45–48]. It discusses how to leverage the data in the DE to evaluate and improve the scientific predictive model.

- Definition of a systematic, automatic, and fair model validation and analysis procedure.
- Development of a data science algorithm based on model validation results to comprehend the behavior of a scientific predictive model.
- How to mitigate ethical problems in the model validation procedure.
- Development of an adaptive sampling algorithm for the design of experiment, and integration of element conservation in the chemical reaction neural network in the context of neural ordinary differential equations.

This thesis concludes with Chapter 8 where the obtained results and future work are discussed.

CHAPTER 2

Related Work

This chapter presents a general overview of the relevant literature to contextualize this thesis's motivations and contributions to the state of the art.

The value of data in several contexts is clear. The current information economy heavily relies on the quantity, quality, and organization of data for future developments [53]. As a result, data markets have become increasingly central. However, the process of systematically getting this value is under investigation. Several approaches are emerging, both technologically and as business models. Data sharing and reuse is an essential part in this context. Unlike data exchange, which only concerns technical aspects, data sharing refers to a broader set of concepts such as access policies, business models, services roles, relationships, and responsibilities [54]. For instance, in data science pipelines, this requirement involves both the publication phase and the preserve destroy phases [55]. In the publication phase, not only making data available on portals, databases, and the like is needed, but also code, workflow management [56], and collecting and aggregating data. In an industrial context, creating data spaces is advocated [57] as "a foundation for the data economy in the European Union". In the Gaia-X initiative, data spaces involve multiple stakeholders, and a set of infrastructural federation services is envisioned to guarantee "identity

and trust," "sovereign data exchange," "federated catalog," and "compliance." In the European Open Science Cloud (EOSC)[1] and Consumer Data Research Centre (CDRC) [58] initiatives the goal is to provide a cloud-based infrastructure for sharing and managing scientific and non-scientific data, enhancing their value, but also a conceptual framework for all needed service, through the ongoing activities of its task forces[2], among which task forces for Authentication and Authorization Infrastructure Architecture (AAI), for Quality Research Software, for Technical Interoperability of Data and Services, and for Long-Term Data Preservation, where needed challenges and new services are being discussed. Data Ecosystems (DEs) are data-sharing platforms defined as "distributed, open, and adaptive information systems with the characteristics of being self-organizing, scalable, and sustainable" [59]. A DE shares data from a *data producer* to a *data consumer* typically through web-services [60], managing the hosted data and increasing their value. Data management is a precondition for the new data science techniques. The data-sharing platform users can achieve results that would not be possible by individual participants [61]. Therefore, it is fundamental to support the data science life cycle with adequate tools and consider it a process composed of multiple phases in which different strategies and tools are needed [62].

DEs, de facto, establish what can be discovered from data [50]. On the one hand, these data-sharing platforms incentivize data reuse, value, and proliferation. On the other open new challenges related to data management [63], such as data quality [59], diversity [40], integration [64], transparency [65]. For instance, if the data are not adequately managed, centralized data management can quickly propagate errors within the system and to the data products [66]. Several studies have listed the design principles of data spaces [67,68], the architectural components [69], particularly in industrial settings for the digitalization of the industries [70], but in practice, each DE has its challenges, based on the application domain, that require specific customization [71]. If not properly addressed, these challenges can completely preclude the adoption of such a promising infrastructure, and as a result, such data-sharing initiatives fail quickly due to the low adoption or interest, such as in the case of the most recent personal data markets [53]. For instance, the medical sector requires a high level of trust and security [72], whereas the main challenges in the energy sector also impose to reach the fulfillment of regulations imposed by the data provider [73]. In general, the *fil rouge* between all DEs in terms of challenges consists

[1] `https://eosc-portal.eu/`
[2] `https://www.eosc.eu/eosc-task-forces`

in building trust between the stakeholders involved [74, 75]. To achieve this, a proper Data Quality (DQ) assessment and higher transparency are examples of possible solutions [65]. Other factors, such as openness and security, contribute to reaching a critical number of users necessary to keep running the platform. On the other hand, pricing and non-interoperable platforms are among the main failure factors [76]. In many cases, the role of a coordinator within the data space enhances the trust between the data consumer and the data provider [77]. Over time, four DE typologies have been defined based on the policy to manage the data, the DE goal definition, the degree of participant interaction, and data exchange within the ecosystem [78]. The first two typologies are Organizational and Distributed DEs. Both have a central control system to fulfill a predefined goal, but the DE participants can operate independently in the first one. In the latter, changes in the DE and pooled resources require participant collaboration. Meanwhile, federated and virtual DEs have no central management authority. In federated DEs, participants interact voluntarily to reach a predefined goal, while in the virtual DE, a coalition of participants can emerge to pool resources to achieve a specific goal. These four types of DE describe the edge cases of authority control, resources management, and participant interactions, but not all scenarios can be restricted to design a DE confined to only one of the previously mentioned DE categories.

One emerging problem of DEs projects is the lack of continuity in time. Some DEs last the duration of a project or an initiative, while others struggle to continue. For instance, previous attempts in the chemical kinetic area are Prime [79] and CloudFlame [80], which started large collections of data of chemical experiments but were discontinued or reduced mainly for the lack of resources. Similarly, many of the data sources for COVID-19 infodemics were discontinued as the pandemic started to be under control, even if not yet terminated; in other cases, the Application Programming Interfaces (APIs) for accessing data may change, or their access policies (such as in the recent changes in policies for Twitter APIs), thus making data inaccessible for some of their previous users building applications on them. Long-lasting initiatives usually have a "single" owner, such as business-oriented companies like Google, or governmental organizations such as National Statistical Offices, or large international organizations, as in the case of data collection for Sustainable Development Goals Indications (SD)[3]. When supported by a data sharing platform, such as in the case of data collaboratives[4], data sharing initiatives may focus on a shorter period of time

[3] https://unstats.un.org/sdgs/dataportal/analytics/DataAvailability
[4] https://datacollaboratives.org/

linked to a given research, even if the platform remains available [81]. Data spaces are multidisciplinary elements, and the business community has already investigated the aspects that threaten their lifetime. Most of these data collaborative are limited, and their impact is contained by the complexity that often a small entity faces in terms of legal, technical, ethical, commercial, and organizational challenges [82]. Fortunately, the promoter of such data collaborative initiatives can potentially circumscribe some of these concerns by developing a more sustainable DE [82]. The most critical factors for data collaboratives are resources, their business model [83], trust, incentives, and data quality [32]. There is, therefore, a gap in the current state of the art for the design and development of the data spaces. At the business level, the main threats are known [84], but there is no in-depth discussion about technological solutions. On the other hand, the technical community potentially has technological competencies for feasible solutions but does not investigate the causes of a short-lifetime data space. This work tries to close this gap by addressing the main business threats and thus presenting technical solutions in terms of data governance [85] aspects. In particular, the proposed solution tries to achieve voluntary data sharing by incentivizing engagement [86, 87] and limiting operational costs [39]. In particular, in the last year, new prominent research results have been about mechanisms to incentivize data sharing in federated learning. These strategies prevent free-riding without the need for any payment mechanism, which would be a deal-breaker in attracting new users into the data collaborative. These mechanisms try to maximize the amount of data generated by each agent [88] or employ a data reward mechanism [89]. Therefore, it is necessary to study the challenges and services needed to create a sustainable DE for a research community.

Such data management systems are fundamental for the Industry 4.0 domain [90], in many industry sectors, e.g., in airline, automotive, chemistry plants, or machine-building industries [90], and scientific sectors, e.g., atmospheric chamber data [91], tsunami-related data [92] and materials science [93]. Supply chains are characterized by lengthy negotiations for agreeing formats of shared data. These processes should ideally be semi-automated, ensuring that they are negotiated, executed, and monitored for contractual and legal compliance, efficiently. As a result, frameworks for data sharing should be characterized by the capabilities of data analysis and the tools for assessing DQ. In general, re-purposing data for analysis and developing models using Artificial Intelligence (AI) technologies require an understanding of the data and their associated characteristics. The data properties are often represented as metadata and, in many cases,

are implicit. Data sharing is also a central element in the modern scholarly debate [94]. In the last years, it is recorded a steady growth in the quantity of shared scientific data. This tendency is recorded across the research groups' disciplines, ages, and geographical locations. People are nowadays more willing to share data to benefit from the citations of their works [95]. Proof of this fact is that generally, a scientific publication that shares data receives about 10% more citations than another work. Another benefit of data sharing in the scientific domain is to reuse resources and increase the dataset's quality. The more users and the more data are present, the more the data are cross-validated. A drawback of data-sharing practices is to agree on the structure, management, and infrastructure to share the data. The more data, the more the representation formats and sources. To address these challenges, it is necessary to define a general architecture for scientific repositories. Such a data management system should address the typical characteristics of data science applications on big data: large volume, acquisition speed, and variety of data [96], but maintaining the Findable, Accessible, Interoperable, and Reusable (FAIR) policy requirements [97] and a quality big data repository [98] since, otherwise, a data of low quality can rapidly spread all over the DE. The recent NIST proposal for a Research Data Framework RDaF[5] provides a well-defined infrastructure for sharing scientific data and tools for the data collection, sharing, visualization, and analysis. It investigates all the phases of a data management process and identifies the users' roles and objectives of the platform. Such domains include the genomic field [99, 100] or the chemical kinetics. Both domains have the problem of sharing heterogeneous data with different quality levels. A DE does not only share data but also services that make the repository informative and capable of extracting valuable knowledge [101]. For instance, in the combustion domain, since the first example of the PrIMe (Process Informatics Model) system [79], these platforms do not exist only as scientific data repositories but offer other domain-related services. PrIMe, in particular, also had the purpose of collecting predictive models and generating them based on specific user requests (e.g., operating conditions), providing services to control the consistency of the experimental data [102, 103], and validate the models [104]. The Bound-to-Bound Data Collaboration (B2BDC) methodology is a part of the PrIMe framework. It is rooted around the concept of consistency, and it is the first methodology that uses data to define constraints to bound a feasible space of variables [105]. As an evolution of PrIMe, CloudFlame[6] was

[5] https://www.nist.gov/programs-projects/research-data-framework-rdaf
[6] https://cloudflame.kaust.edu.sa/

proposed [80, 106]. It offers cloud simulation computing capabilities, a data repository, and a model generation feature. Another framework is ReSpecTh (REaction kinetics, SPEctroscopy and THermochemistry experiments)[7], which contains reaction kinetics, high-resolution molecular spectroscopy, and thermochemistry data [107], and tools to carry on a large number of simulations, validate models, and other functionalities [108]. The diffusion of experimental data facilitated the development of complex models capable of predicting the behavior of thousands of subjects employing tens of thousands of equations. Given the size of these models, it is not easy to keep their development under control and determine which one is the best in some specific circumstances. Thus, several initiatives were undertaken, such as CaRMeN (Catalytic Reaction Mechanism Network) [109], those proposed by West et al. [110], or by Killingsworth et al. [111], which offered tools to check the physical consistency of predictive models, identify errors, and compare their performance. ChemKED[8] is another example of a scientific repository [112]. In all these cases, to the best of my knowledge, no scientific platform entirely embraces the “modern" data-sharing perspective in which the collected big data becomes “smart" data [101], also leveraging the new data science technologies. There is no formalization of the data and the stages involved in the scientific model development process. In addition, it is necessary to identify the properties of the scientific data and the domain requirements to design and implement an open data-sharing platform that offers data and services to support and improve the end-to-end process, guaranteeing, for instance, certain data quality levels, transparency, interoperability, and scalability.

A central activity in the model development process is model validation. In the case of scientific predictive models, typically, the model predictions are validated against the experimental data. For decades, the methods to establish whether or not the model predictions are congruent with experiments have been of great interest in combustion research [113]. Some of these methodologies take only into account the distance between measurements and predictions, with metrics such as mean square error [114–116], R^2 [117, 118], or customization of them based on the application, referred as objective error function [119]; others also consider the dissimilarities and similarities among the shapes of the experimental and simulated data curves [120, 121]. However, a systematic approach to compare and analyze different numerical model validations on very large quantities of data is needed, in addition to new ways of analyzing in-depth critical cases when

[7] `http://respecth.chem.elte.hu/`
[8] `http://www.chemked.com/`

identified, exploiting all available data. Therefore, an information system is essential to manage, interpret correctly, and analyze scientific big data automatically.

The word *uncertainty* is used to generally describe a lack of knowledge about the explanation and description of phenomena [122]. Aleatoric and epistemic are two macro-categories of uncertainty [123]. Aleatoric uncertainty concerns the intrinsic randomness of the observation of phenomena. For instance, the imprecision committed in the measurement of experiments or the fuzziness of an image. On the other hand, epistemic uncertainty is related to the representation of a complex domain in a predictive model with a reduced number of variables. Such simplifications lead to uncertainty in the predicted values [123]. Uncertainty can be classified in more specific categories [124]. For instance, data uncertainty, due to the finite precision of instruments to represent a continuous world, leads to random errors or does not account for other uncertain sources, such as the instrument drifts that, instead, lead to systematic errors, as well as sampling errors [125]. Uncertainty, if provided, has to be properly managed [126] and algorithms [127, 128]. There are two ways to represent the uncertainty in the data [129]. The first one represents the uncertainty as a probability distribution of probabilities of the available data rather than deterministic facts. Otherwise, it can be represented as metadata that specifies statistical information, such as the average and standard deviation.

When it is necessary to develop a predictive model from experimental data, the uncertainty of the data has a central role in the development process. To mitigate the Garbage In - Garbage Out (GIGO) effects [130, 131], experimental uncertainty is fundamental to assess the quality and reliability of the data. If so, the data can be used to guide the model development, but often, the uncertainty is not reported [132]. The lack of such investigation regarding the uncertainty of the data in scientific fields is mainly due to two elements: either the impossibility of replicating the experiment in the same conditions or the high data collection and generation cost [132].

To correctly develop a of predictive model is necessary to estimate the missing experimental uncertainty [133]. There are two possible methodologies to quantify the uncertainty of the measurements [134]. The most accurate one is if it is possible to replicate the measurements inexpensively to quantify the experimental uncertainty with an extensive experimental campaign [135], where a sufficient number of measurements is needed to estimate uncertainty accurately [136]. The average of the measurements is the best estimate for the value to be reported, and the uncertainty is equal to the standard deviation [134]. The second methodology leverages the Taylor

series expansion when it is not possible to measure a quantity directly [137]. Another novel approach to estimating data uncertainty of physical phenomena when it is not possible to replicate an experiment leverages the fact that the dependent variable changes smoothly when each independent variable changes a little while others are kept constant. This assumption allows the use of regression models, and the model residuals can be used to estimate the uncertainty of the dependent variable [132]. Due to the cost of the scientific data, The major challenge of such an approach is related to the limited availability of data in similar conditions (i.e., independent variables). Moreover, in complex domains, the general assumption that a slight change in the dependent variables corresponds to a small variation in the independent one is not always true or easy to assess how much is a little change. Finally, a naive approach is to use default domain values [115] to complete the missing experimental uncertainty. However, the uncertainty could differ significantly from the suggested default value [36].

Metadata is known to be used to model the DQ [138]. Since, uncertainty can be seen as another metadata of the experiments, it can be linked to DQ [139]. Generally speaking, data, DQ , and uncertainty are related by data profiling [140]: current DQ reports are imprecise since they lack complete descriptions of data uncertainty [139]. Moreover, DQ indicators, independent from the accuracy of the procedure to account for the data uncertainty, they do not account for the adequacy of the data for a given goal [141]. Metadata can be used to construct an ontology of the domain included in data [142, 143], and knowledge graphs [144] can define ontologies for scientific data domain [145]. Therefore it is possible to leverage their structure to profile the data and measure the DQ effectively [146].

Knowledge Graph Embedding (KGE) is a field of Machine Learning (ML) that learns how to represent a Knowledge Graph (KG) in a low-dimensional space. Such embedded representation can be used by some ML tasks such as "link prediction", to infer missing relationships between two existing entities in the graph [147, 148], focusing on experiments and their uncertainty value. Embedding represents a complex entity in a lower-dimensional space such that entities with similar semantic meanings have close embeddings.

Predictive models are developed following a data-driven black-box or white-box approach. However, regardless of the methodology, the time required for prediction or simulation can be particularly computationally expensive and time-consuming. In the case of chemical kinetics, the prediction time given a white-box model can last even weeks. The development of an approximation of predictive models can overcome this limi-

tation. In recent years, various adaptive sampling algorithms have been studied to optimize the process of developing metamodels that approximate a function in a complex domain. These algorithms, at each iteration, refine the sampling strategy balancing between exploration of the unknown domain areas of the function response surface and exploitation of the information obtained from previous sampling steps. The choice of an appropriate sampling algorithm hinges on the specific application. Several interconnected factors influence this decision. Firstly, the user goal. It specifies whether the interest is in generating a metamodel with general approximation capabilities or focusing on optimizing the learning of the prominent features of a complex function. Secondly, the known characteristics of the function. These characteristics, such as patterns or strong non-linearity, can be leveraged to select the most effective adaptive method. Lastly, the properties of the experimental design. Such aspects regard the initial sample size, the problem's dimensionality, and the risk of clustering. For instance, exploration-focused techniques are preferred for smaller initial sample sizes, while lower complexity methods are preferred to save computational time in high-dimensional spaces. Similarly, to develop a scientific black box using, for instance, Neural Network (NN) requires a significant amount of data that is challenging to collect from experiments or generate from simulations in scientific domains. Therefore, adaptive sampling algorithms can be used to select the smallest and most informative set of data, therefore employed in the Design of Experiment (DOE) phase. A recent review paper offers a comprehensive comparative analysis of different adaptive sampling algorithms [149]. When the user goal is global metamodeling, adaptive sampling techniques with a higher exploration component, such as MASA [150], MEPE [151], and MIPT [152], are found to be more proficient. However, WAE [153] and EI [154] demonstrate the worst performances. For optimization, MEPE [151] and EI [154] outperform other strategies, although EI [154] is heavily dependent on the initial sample size [149]. Regarding known response surface characteristics, most adaptive sampling methods perform well for regular patterns, with minor differences in performance. In the case of irregular patterns, adaptive methods that emphasize exploration show better approximation abilities. Concerning the properties of the initial experimental design, the paper discusses the importance of initial sample size, the parametric dimension of the problem, and the risk of clustering. Techniques that emphasize exploration, such as MASA [150], MEPE [151], MIPT [152], and SFCVT [155] are favored for smaller initial designs. Lower complexity techniques are preferred to save computational time in cases of high para-

metric space dimension. Additionally, the review considers supplementary and miscellaneous criteria, such as versatility, computational costs, coding complexity, and optimization problems given by cost functions. Across all the examined criteria, MEPE [151] offers the most comprehensive performance. This method is particularly recommended when there is no prior knowledge about function characteristics. EIGF [156] and MIPT [152], an exploration-based technique, provide reliable results and require less development and user knowledge. However, there is no one-size-fits-all solution. Achieving optimal results necessitates striking a balance among competing factors, highlighting the necessity to introduce a novel algorithm that addresses these considerations more effectively.

The scientific domain of interest for this thesis, chemical engineering, has demonstrated the successful adoption of data science [20, 116] or, more in general, computer science such as Principal Component Analysis (PCA) or KG approaches to extract new knowledge from data [157, 158], or inside an optimization procedure of existing models [119]. Furthermore, since many black-box ML applications are spreading in this research area [159–164], it is important to apply and adapt to the chemical engineering domain the existing expertise in the computer science community to avoid well-known issues, like model bias [165], and to gain knowledge from the increasing amount of experiments, simulations, and predictive models. Few of these black-box MLs models have included well-known physical laws in the model discovery procedure to be leveraged during the training. Most of them use this information to penalize wrong predictions during the training by adding penalization terms in the loss function, while others inherently impose to satisfy physical constraints [166, 167]. However, it is still an open issue to interpret black-box physical models. Chemical Reaction Neural Network (CRNN) and its evolution [168], demonstrates that it is possible to combine the generalization capabilities of NN with physical constraint while learning an interpretable physical model. Applications of such framework are successfully applied to optimize a pre-existing kinetic model [169] for battery thermal stability analysis [170] and biomass pyrolysis cases [171]. It is still to investigate the application of such a framework for other kinetic models, such as Hydrogen.

CHAPTER 3

Scenario

For this thesis, having prior knowledge of chemical engineering is unnecessary. However, this section briefly describes the domain scenario without detailed technical explanations and limits the description to the concepts needed to better understand the challenges, requirements, proposed methodology, and results. Over the past few decades, the progress in computing power, the availability of more and more data, and the tendency to share information boosted the development of many research and industrial areas [172]. The availability of predictive models to forecast a system state brought new insights into comprehending the phenomena, industrial applications, and social benefits in many different sectors, from engineering to social science [2]. The applicative scenario of this thesis is in chemical engineering, more precisely, chemical kinetics. Chemical kinetics studies how experimental conditions influence the speed of a chemical reaction [30]. The experiments are then used to derive information about the reaction's mechanism and transition states, with the final goal of the construction of predictive models that can describe the evolution of a system [30]. Due to a large amount of available data, the generation procedure of complex predictive models is changing from an approach solely based on first principles to data-driven methodologies [17, 173, 174]. As

these phases require a considerable effort, and due to the complexity and the many possibilities to model a domain, the “many-data many-models” problem originated [175]: many models are available to predict the same subject (i.e., the quantity or property of interest), but they differ in the number and form of mathematical equations representing the phenomena or in the selection of parameters [175, 176]. These degrees of freedom and the “many-data” led to the development of many models of various complexity from different research groups concerning the same subject but based on a different subset of experiments. The result was the generation of inconsistent and not general models [145]. In addition, a manual evaluation of model quality through comparison with experimental data and a univocal, quantitative ranking of the results are not straightforward operations [109]. The diffusion of experimental data facilitated the development of complex models capable of predicting the behavior of thousands of subjects employing tens of thousands of equations [17]. Therefore, there is the need to organize the available information and conceptualize the problem in terms of big data, automate the model validation and analysis procedures, and extract knowledge from the data to speed up the development process while reducing error-prone tasks [109]. For these reasons, chemical-physical predictive models often are data-driven models [120]. Over time, there were several initiatives aimed at collecting experimental data in a so-called Data Ecosystem (DE). Their typical challenges are the involvement of the scientific community in data sharing, providing services to users, and the standardization of data representation in agreed formats [177].

Figure 3.1 describes the business process of the predictive model development loop, involving four types of scientific data, experiments, models, simulations, and analysis results, which will be described more precisely in Section 4.1. This loop includes four main stages that have been identified within this thesis, and, to the best of my knowledge, no prior formalization has been done in the literature. This process is the starting point from which the process needs to be improved, for example, including new phases, automating the existing ones, and defining roles and responsibilities. It begins with the collection of experiments. Based on this new information, a new model or a previous version is generated or improved to represent the new experimental data, if necessary. Later, the predictive model simulates the same domain condition of all or a subset of the available experiments. Finally, the analysis starts. Experiments, i.e., the “ground truth" values, are used to compare the model predictions, i.e., the simulations. This particular type of analysis is called *model validation*. The analysis results have synthetic insights about the model performance. This information is then used

in the model development phase to improve the current model. The model development cycle is repeated until the analysis of the results is considered satisfactory or no more new experimental data are available.

A DE to support this process needs to offer both data repository capabilities and services on the data to speed up and improve the process. The DE users belong to one or more scientific organizations or groups, such as a department, research group, or university. In our scenario, Politecnico di Milano, with the CRECK modeling group [178], is the founding partner of this initiative, supported by the RAISE group for the information system part [179]. This peculiarity defines the interdisciplinarity of this thesis.

The data, such as the experimental observation of a chemical system and the predictive models used for this thesis, are from the combustion kinetics domain. This research area is currently a central topic for the energy transition to investigate the capabilities of carbon-free fuels, like hydrogen or ammonia, to improve the efficiency of current fuels or to develop new ones [18]. Experiments are experimental observations of a phenomenon in a physical-chemical system. Different kinds of experiments are, for instance, *ignition delay time* and *concentration-time profile*, and are done in experimental facilities, “reactors", such as a *shock tube* or a *plug flow*. These experiments study how the initial composition and the physical properties of the experiments, i.e., the initial mixture of species contained in a reactor in a given environmental condition, such as temperature, pressure, and volume, evolves until the reaction process is completed. ReSpecTh [107] defines the ontology of combustion experiments. According to the ReSpecTh ontology, to adequately describe an experiment, it is necessary to define the source of the data and the initial environmental conditions of the experiments that are considered the metadata of an experiment. These metadata are fundamental since, for instance, the evolution of a chemical system highly depends on the initial environmental condition and mixture of species. Instead, the data are the variation of the measured chemical-physical properties, i.e., the subject of the experimental investigation, during the reaction process.

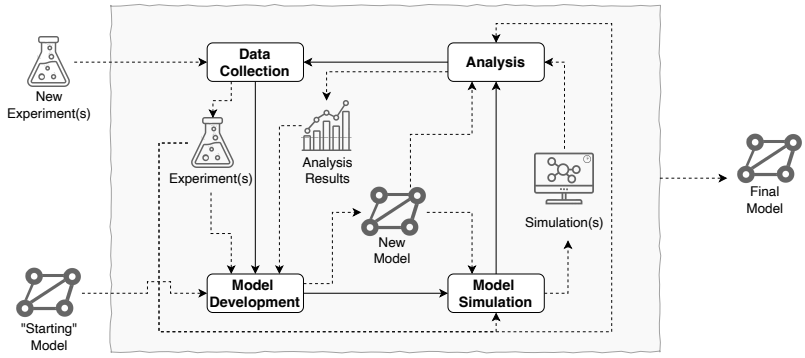


Figure 3.1: *Business process of the predictive model development loop.*

CHAPTER 4

Data Ecosystems for Scientific Data

Data Ecosystems (DEs) collect, manage, and analyze large volumes of heterogeneous data. As introduced in Chapter 2, gathering data in the same platform facilitates data reuse and sharing, as well as the discovery of latent insights by applying data science techniques. The design of such a platform differs from the purpose of the DE itself. As previously explained in Chapter 1 and in Chapter 3, this work studies the design and application of a DE to support and improve the development process of a scientific predictive model. For this purpose, a set of services on the data hosted in DE will be necessary to be provided to support such activity. The detailed development of the DE depends on the specific application domain. Generally, two elements mainly affect the design of such DE.

The former is the data to be hosted [63]. It is essential to identify the distinct typologies of data used during the development process of scientific predictive models. Data may differ in sources, representation formats, and data quality levels, making conventional approaches for the DE design not perfectly suitable and demand particular attention [63]. For this reason, after identifying the main different types of scientific data in Section 4.1, Section 4.2 investigates the scientific data properties and what makes the data integration in the DE challenging.

The latter aspect regards the domain requisites by stakeholders and the specific circumstances for which the DE is conceived [63]. For instance, one emerging problem of DE projects is the lack of continuity in time [32]. Some DEs last the duration of a project or an initiative, while others struggle to continue. Consequently, Section 4.3 first introduces and examines the business motivation and circumstances that drive and challenge the adoption of a DE in a scientific domain. Then, it formalized the scientific domain requirements expressed by the scientific community that must also be considered while designing a DE for scientific applications.

Finally, Section 4.4 summarizes the challenges that must be addressed to design and develop a DE for a scientific domain. Therefore, this chapter addresses the first research question introduced in Section 1.3.

4.1 Scientific Data

This thesis proposes to classify *scientific data* in four different types of data: experiments, predictive models, simulations, and analysis results. These data types are collected, used, or produced within a cyclic development process aiming to deliver a predictive model, as introduced in Chapter 3.

Figure 4.1 presents a high-level and non-domain-specific class diagram of scientific data with their main attributes (or metadata) and their relationships. The detailed class diagram can involve additional classes, attributes, and relationships in a real-world scenario in a specific domain. For instance, Figure 4.2 depicts the class diagram of the different scientific data for this work's scenario.

4.1.1 Experiment

The term *experiment* represents data from an experimental campaign carried out by an experimenter. According to the ReSpecTh ontology [107], an experiment combines chemical-physical measurements and associated metadata. The measurements account for quantifying a property under investigation, while the metadata provides essential details regarding how, when, and by whom the measurement is carried out. Examples of this information are the technical instruments used, the environmental setting, the unit of measurement, the identification of the property of interest, the experimental procedure, and so on. In essence, the experiment metadata describes the experimental condition (or setting) of an experimental campaign. On the other hand, the measurement is the numeric value detected by the instruments and thus reported (*Reported value(s)*). Due to intrinsic measurement errors, usually, an experimenter carries out an experimental

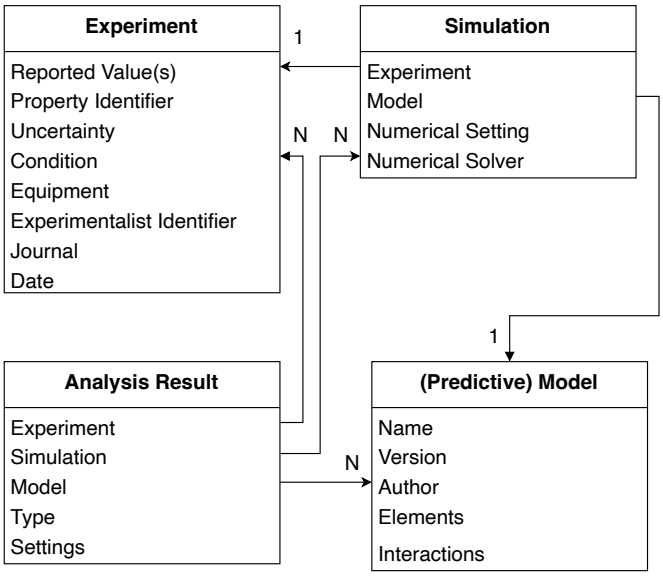


Figure 4.1: *General class diagram for scientific data.*

campaign rather than single experiments, i.e., the same experiment is repeated multiple times in the same experimental *condition*, in order to be able to quantify and report the *uncertainty* on the measurements. The experimental condition expressed by the metadata does not vary between the multiple measurements. Usually, the experiments are published on a *date* with an associate publication in a *journal*. In most cases, the digital size of the data necessary to represent the essential digital information, i.e., the *experimental data* that consists of the reported value and metadata, is quite tiny. It is rarely bigger than 10MB, and often it is less than 1MB, even if the entire material needed to derive the reported value can have a different order of magnitude in size. Therefore, with a moderate storage cost, storing a large amount of experimental data is generally possible.

4.1.2 Predictive Model

A scientific predictive model is the business driver of a DE for scientific data [180]. The final purpose is to deliver an accurate model to predict unknown outcomes. Nowadays, a popular type of predictive model is Neural Networks (NNs). NNs are black-box methods by design. On the other hand, in the scientific domain, predictive models usually embed chemical-physical laws into equations such as chemical reactions [17], which cannot be violated. Most of the time, chemical-physical equations can be trans-

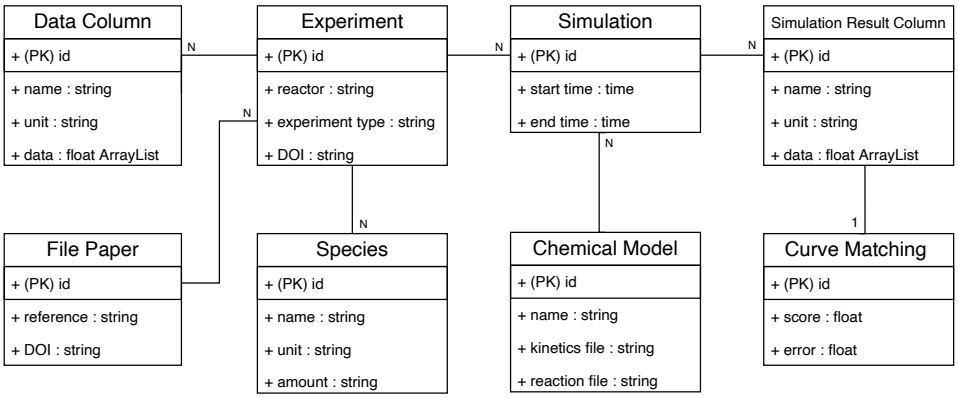


Figure 4.2: *Class Diagram for SciExpeM DE: Representing Experiments, Simulations, Models, and Analyses (with (PK) denoting primary key). Some entities and attributes are omitted.*

lated into a set of interpretable differential equations [30]. Thus, these are white-box methods. Black-box methodologies, such as Physically Informed Neural Network (PINN), can also be employed to develop predictive models for scientific domains. PINNs incorporate chemical-physical laws in the loss function to learn a set of parameters (NN weights) [181], but they are not interpretable. PINN, and in general NN, require much more data for the training and parameters to represent the domain, and they usually do not generalize as well as the white-box methodology [181]. In scientific domains, in the case of white-box models, which elements and equations are included and how they are included in a scientific model is a design choice of the researchers, and it is usually referred to as model parameters [175]. In general, the larger the number of equations in a scientific model, the more complex and accurate it is to resolve it. Simplifying, in the general case, a scientific model tries to predict how a chemical-physical system evolves starting from a particular initial condition, solving a set of equations that encode the elements and their interactions in a domain that the model designer decided to represent. Like neural networks, scientific models are also defined as data-driven since real-world observations, i.e., experiments, are used to validate and improve the predictive models.

4.1.3 Simulation

A predictive model, given a set of initial conditions, can forecast the future state of the represented system. The predicted state is referred to as a

simulation, which represents the solution to the model equations obtained through a numerical solver. Numerical solvers are generally very complex, thus time-consuming, since they need to resolve, for instance, differential equations and, in general, need numerical tweaks to address the problem correctly [182]. A wrong or inappropriate configuration of the numerical settings can lead to incorrect results, even if the underlying model is accurate. Additionally, improper numerical parameter settings may lead to excessively long computational times and, in some cases, to failures to terminate the computation. The output file size of a simulation can vary based on the grain of the numerical settings and their complexity. For instance, their file size can range from less than a MB to several dozens of MBs.

4.1.4 Analysis Result

The analysis results in the process of developing a predictive model serve the purpose of generating synthetic information on scientific data. More specifically, it is intended to provide aggregate information on the model's predictive capabilities in different domain settings. Which metric or procedure to employ during the analysis is a parameter of this type of data.

4.2 Properties

This section discusses six properties of scientific data that make their management and integration in a DE challenging. Table 4.1 qualitatively summarizes the high (H), medium (M), or low (L) impact or relevance of property on a specific type of scientific data, as described in Section 4.1.

4.2.1 Low Volume - High Cost

Scientific data, unlike other types of data, such as social media, are less available [25]. The collection of experimental data involves on-field measurements using expensive equipment and materials, making it costly in terms of both financial resources and time [26]. Consequently, experiments are often unique and not easily replicable. Similarly, developing predictive models is a complex process that demands extensive expertise, particularly in white-box approaches [30], and extensive computational resources and data for building models in black-box approaches, resulting in only a limited number of models being available for certain scientific domains. Simulations are also available in a limited quantity since they are very pricey in terms of computational resources needed and, in some cases, space to store them [182]. Consequently, the results of analyses based on the other three

Property	*Experiment*	*Predictive Model*	*Simulation*	*Analysis Result*
Low-Volume High-Cost	H	H/M	H/M	L
Uncertainty	H	H	M	M
Accuracy Consistency	H	M	L	L
Heterogeneity	H/M	M	H/M	L
Completeness	M	L	M/L	M
Reproducibility Transparency	H	M	M	H

Table 4.1: *Qualitative impact of a scientific data property on the corresponding scientific data type - (H) high, (M) medium, (L) low.*

types of scientific data are limited, too. Analyses are generally relatively inexpensive to compute. Although scientific data volume is lower than other types of data, manual management is still unfeasible and prone to human error. The low volume and high cost of scientific data highlight the need for a data management system, such as a DE, in the various scientific domains to promote the reuse of all types of data and related development and analysis services.

4.2.2 Uncertainty

Uncertainty can be classified into two macro-categories: epistemic and aleatoric [123]. Experiments are real-world measurements, and they are intrinsically affected by aleatoric uncertainty. Repeating the same experimental measurement helps mitigate this issue, quantifying the uncertainty. The reported value for an experiment corresponds to the mean value of the measurements, and the standard deviation corresponds to the uncertainty [135]. The source of uncertainty in the experiment is not only due to measurement error but also a set of contributing uncertainty causes. For instance, another source of uncertainty for the experiments is the digitalization of plots from physical documents, such as published papers and reports, to extract the measurement values. Models, on the other hand, are mainly affected by epistemic uncertainty. A model is an approxima-

tion of a real-world system, inherently introducing errors. Simulations are, most of the time, deterministic [182]. Repeating the exact simulation of an experiment with the same model leads to the same result. However, numerical errors can also affect the uncertainty of the model's predictions [182]. Analyses are generally not uncertain, even though they are affected by the propagation of uncertainty from the models, simulations, and experiments. The uncertainties present in these underlying components can influence the overall uncertainty in the analysis results.

4.2.3 Accuracy & Consistency

Experimental observations of the same chemical system should be close to the (unknowable) ground truth and consistent with each other. If multiple experimental measurements are available from different sources regarding the same experimental conditions, all the reported values should be consistent, also accounting for their uncertainty. In other words, the more the reported values are accurate, the lower the experiment's uncertainty is, the easier it is to detect inconsistencies. Nevertheless, in reality, it is hard to evaluate the consistency of the (many) experiments without uncertainty. Models represent the interaction of the system elements. However, when new elements are being investigated, they may not be standardized in the representation [110]. For instance, models can represent different entities using the same element name. As a result, it is not easy to compare the simulation results consistently. Numerical solvers differ mainly for numerical implementation choices such as the number of digits or the employment of a particular library [182]. Thus, giving the same model and conditions for forecasting can lead to different numerical solutions (simulations). Usually, the difference is marginal. Finally, consistency is almost guaranteed concerning analysis data, assuming the analysis procedure is well-detailed and fixed on the same dataset.

4.2.4 Heterogeneity

Scientific data exhibit heterogeneity from three distinct perspectives: type, source, and format [29]. Since many scientific domains have been active for several decades, the sources, resolution, and methodologies for collecting and doing experiments and models have evolved. At the same time, it is likely not to have a standardization or an ontology by the scientific communities on the representation format [180]. However, there are (sometimes also multiple) de-facto representation formats for all types of scientific data.

4.2.5 Completeness

Experiments are usually produced and collected over several decades. As the way of collecting them and scientific findings change over time, some additional information may become essential to include among the experiment metadata; however, some (old) experimental data may have incomplete information when more recent metadata are considered. Models are incomplete by definition since they simplify a real-world system. For instance, it may not include all the elements of the domain or all the interactions. As previously stated, the effects of such decisions are reflected in the model prediction accuracy and uncertainty. Simulations report all the elements and interactions described in a model, but are quantified along discrete and, thus, not continuous, dimensions.

4.2.6 Reproducibility/Transparency

Experiments are challenging to replicate since it is practically impossible to reproduce the exact initial conditions of an experimental setting [183]. Models and simulations, instead, if adequately documented, are easily reproducible. It is fundamental that models disambiguate the meaning of represented elements [110] and, concerning the simulations, the numerical settings [182]. Regardless of the methodology to derive a predictive model, white or black box, explaining the simulation results could be difficult [184]. Analysis results must be transparent about the analysis process and the computational steps to avoid inappropriate conclusions [65].

4.3 Requirements

Chapter 2 introduces that both academic and industry research departments, over time, have established principles, methodologies, and approaches to address the initial challenges in designing and developing a DE. However, there is ample room for improvement when it comes to thinking about the countermeasures to sustain these projects over the long term. These platforms often experience a significant rate of early-stage failure, primarily due to low user engagement and high costs. Therefore, it is essential to consider these factors, especially during the design and initial deployment phases. The factors contributing to this phenomenon, as introduced in Chapter 2, are already being investigated at the business level. However, the technical aspects of incorporating these findings into the design phase and the technology require further discussion. Therefore, this section first identifies the main reasons that threaten the long-term use and life of a DE

while also formalizing the requirements in terms of costs and user engagement. Later, Chapter 5 will discuss the mitigations at the design level and the potential technological solutions.

Three main macro-phases depict the life cycle of a DEs [39]. The first one, the *Initial Creation*, creates a data collaborative environment. During this stage, data is collected and curated, often in the context of specific research projects, and is stored and shared with interested parties. In the second *Community Building* phase, the data collection process is more systematic, and community computing services may be introduced to facilitate data analysis. Finally, the third and last *Maturity* phase sees new DE services development. Consequently, maintenance support becomes crucial, and effective user engagement and management are essential for ensuring the continuity of services. As shown in Table 4.3, these three macro-phases may entail varying costs, frequencies, and partnership models.

To thoroughly analyze various scenarios in detail, Table 4.2 provides an overview of several analysis dimensions. These dimensions represent characteristics that impact sustainability and governance and are considered in relation to the maturity of platforms in the later macro-phases as described below. These dimensions differentiate between the *roles* played by involved stakeholders, which, in some cases, may be undertaken by the same individuals or users. Typically, during the *Initial Creation* phase, these roles are fulfilled by a single researcher (e.g., a Ph.D. candidate), a team, or a project. However, multiple stakeholders usually become involved as the DE evolves. Nonetheless, this division of responsibilities and participation in the platform also influences the project's sustainability. The table categorizes both *Sustainability Challenges* and *Governance Challenges* along with some *Technical Mitigations*, all of which are discussed further in the following sections.

It is not unlikely that DEs fall into disuse after the *Initial Creation* phase that is supported by a starting investment and enthusiasm. In the long term, one of the main reasons for this decline is the absence of a business sustainability plan. In fact, sustainability considerations often do not enter, in some form, into the technical requirements of the stakeholders during the first phases of DE design and development. This thesis proposes the trust-user-data framework to explain this pitfall, illustrated in Figure 4.3. The framework foresees three elements that keep running a DE: user, data, and trust. If any of these elements are missing, it can trigger a vicious cycle, leading to the failure of the others and potentially the demise of the DE itself. With more data, the trust in the DE increases (e.g., cross-data validation), and therefore more users (new and active) are attracted to use the

Role	Sustainability Challenge	Governance Challenge	Technical Mitigation
Producer Consumer Manager	Engagement Confidentiality Costs Duration in Time Source Availability (eg Twitter)	Responsibility Funding Trust Data Quality Policies	Data Architectures (virtually, centralized, federated, P2P, etc.) Computational Infrastructures Interoperability Services Data Lakes, Catalogues

Table 4.2: *An overview of key dimensions of DE. The table provides a concise summary of the roles, sustainability and governance challenges, as well as the technical mitigation strategies discussed in this paper.*

platform. It is also true vice versa (red-dashed arrows in Figure 4.3): the more the number of users increases, the more the platform (and its data) is tested, and thus it is more trustworthy. A trustworthy DE incentivizes the existing user to share more data. With more data, more (new and active) users are attracted to leverage the advantages of the platform. On the other hand, an increased number of stakeholders raises additional issues that require specific management services for identity and trust management [57]. From the platform manager's perspective, user, trust, and data determine the sustainability of the DE, which can be investigated into two macro business aspects: cost and engagement, as shown in Table 4.2. The duration in time of a data-sharing platform can be seen as a consequence of the high costs, low engagement, or discontinued source available data.

Sustainability poses a significant challenge, especially when resources are scarce, and a DE relies on funding from a single entity rather than a consortium representing the community it aims to serve. This study focuses on this (worst-case) scenario in which these conditions prevail. This scenario is not uncommon, as there are numerous cutting-edge domains in both research and industry where the DE serves as pioneering technology, requiring acceptance from an established community, if present. In practical terms, the most common scenario arises when an organization, which we will henceforth refer to as the "founder" (often a university), seeks to implement this technology to advance research in a new or niche field by facilitating data sharing. Consequently, in this context, the *user* is both *data producer* and *data consumer*, affiliated with one or more organizations.

The founder could initiate a crowdfunding campaign targeting potential future users. However, requesting a usage fee for the DE does not encour-

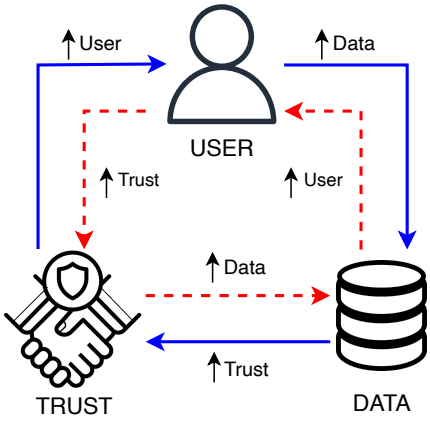


Figure 4.3: *New and active users trust the platform, and the data, quantity, quality, and novelty of the data are the fundamental elements that keep running a Data Ecosystem*

age its adoption and may deter the most skeptical organizations. In the context of a DE, a sustainable business model should pivot toward a different currency: data. In sectors where this technology has yet to be adopted and finding partners is challenging, promoting and launching the initiative implies that the available data is precious and scarce. Consequently, the DE's role in facilitating data sharing becomes even more critical, justifying the initial solo investment. Over time, the value other users generate through shared data will cover the initial investment costs.

4.3.1 Cost

Starting and maintaining a data-sharing platform entails various types of expenses. In our specific scenario (as detailed in Chapter 3), a small organization, often referred to as the funding partner, founder, initiator, or promoter, operates within a scientific research field focused on developing a predictive model. While recognizing the value of data sharing, this organization typically operates with limited resources. Consequently, as the platform expands to include more organizations or users, it necessitates increased resource allocation to accommodate growing demand. However, if the platform relies on limited resources and makes non-sustainable design choices, it may result in unreliable services. Consequently, the platform's appeal and trustworthiness may deteriorate over time. Making judicious design choices that enhance the platform's cost sustainability is critical to address these challenges.

Based on the experience of building a community for sharing high-

quality data among researchers in the chemical kinetic domain [36] and leveraging citizen science in social media analysis [42], Table 4.3 provides a comprehensive overview of ten cost-related aspects. This includes a qualitative assessment of the costs, their frequency, the entities responsible for covering these expenses, and the specific project phases in which these costs should be sustained.

The founder takes on the initial expenses associated with platform development and data collection. The platform's viability hinges on reaching a critical mass of data, as without it, the platform lacks purpose, and users lack motivation to engage with it. In most cases, a DE should host thousands of data whose typical file size is in the order of magnitude of a few MB. In the current scenario discussed in this thesis, data is a scarce resource, and the DE primarily serves the purpose of data collection and information sharing. It is improbable that the DE would need to store millions of large video files, for instance. As a result, the cost of storing data is relatively limited, especially considering that the average price per gigabyte (GB) in 2023 is approximately $0.015\ USD$ [1]. A single entity, such as the funding partner, can generally cover such a cost.

Similar reasoning for the computational resources needed to provide the DE services. The services enhance the user experience, and the ratio between service requests and data presented in the database is very high, so they are very frequently requested. There are many kinds of services. Some of them are computationally inexpensive, and others are very costly. Without fees or large funding, it is impractical for a single organization to bear such expenses in this setting.

Together with the computational cost, the DE users should be in charge of the data collection and their insertion into the platform. This could happen quite often with a non-negligible effort required. However, based on the availability of ad-hoc tools in some scientific domains, some costly and time-consuming data management aspects of the repository, such as data collection, can be automated or semi-automated. For instance, ChemDataExtractor [185] is an automatic tool to extract chemical information from the scientific literature. Nevertheless, in the end, the users' purpose and responsibility (as data providers and consumers) is to provide and consume the data. Similarly, if an organization wants to implement other services, it is free to do so since the platform's source code should be open source, and the platform manager should only integrate it after verification. The effort, in this case, is high but only happens occasionally.

[1] `https://diskprices.com/`

Item	Cost	Frequency	Responsible	Phase
Web Server	++	one-time	Promoter	Initial
Initial Development	+++	one-time	Promoter	Initial
Initial Collection	+	one-time	Promoter	Initial
Data Store	+	+	Promoter	Initial
Data Collection	++	++	User	$\geq$ Community Building
Data Insertion	+	++	User	$\geq$ Community Building
Computing	+++	+++	User	$\geq$ Community Building
New Services	+++	+	User	Maturity
Maintenance	+	+	Manager	$\geq$ Community Building
User/Data Retention & Acquisition	++	++	Manager	$\geq$ Community Building

Table 4.3: *Technical costs that are needed to face to create and maintain a data ecosystem.*

4.3.2 Engagement

User engagement is related to the number of active and new users that utilize the DE. Following Figure 4.3, to increase engagement, it is necessary to enhance the trust in the platform and in the data while increasing the number of new data, i.e., incentivizing the sharing of the data among the platform users.

Scientific domains are highly active research areas with a long history of studies. Over the years, the research and scientific processes have undergone continuous changes. Nowadays, the workflow is consolidated, but recent technological advancements present new opportunities to improve some aspects of the research process. New technologies such as machine learning promise to enhance the comprehension of phenomena [186], while data management systems are fundamental to automating and managing an increasing amount of data [50].

However, even if the new technologies are promising, changing the workflow that has guaranteed continuity of results over the years is problematic. Moreover, these technologies are often distant from a scientific community's expertise and, thus, harder to understand and trust. In the end, proposing new technology, such as the DE, requires keeping the final user and the community involved. Having data and services in the same platform is a game changer because it incentivizes the user to stay inside the platform and switch its usual workflow to a platform-oriented daily workflow. A hybrid configuration in which the data are hosted by the DE, but the users have their own workflow using different technologies for apply-

ing services on the data could not be a long-term solution. It would imply splitting and doubling the workload among different technologies since it is required to move the data back and forth frequently, which might be critical in case of large volumes. On the other hand, staying inside the platform will naturally bring more data, requiring less organization effort, but the DE must integrate all the usual services and sustain the computational cost, as discussed previously.

In a DE where the central focus is data sharing, if users stop participating, it can lead to a decline in data sharing, ultimately discouraging others from taking part. As said previously, in the scenario of this thesis, the data consumer and the provider are not separate entities, but the entities have both roles. In this situation, it is harder to incentivize data sharing because all the users are interested in the platform for the information that they can get from it. There are also other edge situations in which users are more willing to use the DE to promote their data. However, this situation is rare because data generation is generally costly; thus, the ratio between uploading and downloading data is very low. Therefore, a DE can easily be in a deadlock situation in which users are not incentivized to share their precious data if others share a few data.

Two factors require attention to break this negative vicious cycle. First, providing a system that offers all the traditional functionalities and the new ones in a user-friendly way. If some of the traditional functionalities are not present or not working properly, then the final user will not be willing to use multiple methodologies, thereby disrupting their workflow. Although it is acknowledged that software or information systems have a higher probability of failure in the early stages, on the other hand, as user numbers and usage increase, reliability improves. Thus, an increasing number of users translates into more data and greater trust in the shared data and platform. Users employ data in their daily work, and many of these applications are data-driven; thus, data directly impacts their work results. Tracing the origin and keeping control of the quality of the data is, therefore, mandatory to achieve trust in the data, as well as if more users use the data and the platform services, more trust in the DE is generated.

Second, a large initial investment is needed at the beginning to collect shareable data and to implement the DE services and functionalities that can attract users to start using the system and contribute additional data to increase the amount of shared data. Also, in this case, user involvement is important, both for data collection and for defining and testing the needed functionalities, as well as improving them.

In this thesis scenario, encountering data protected by confidentiality is

highly likely. The development cost and potential applications of new experiments and models render them exceptionally valuable. Both companies and academic institutions often hesitate to release their work immediately upon its development. These data provide a technological advance for their upcoming generation of products. Typically, such data only become accessible after publication in journals or through patents. Consequently, for a DE aiming to host scientific data, ensuring confidentiality for data that has not yet been published is critical. To achieve this, two critical factors need to be enforced. First, users must trust the DE to guarantee the confidentiality of their data. Second, it is essential that all data, whether subject to confidentiality restrictions or open access, coexist within the same platform but under appropriate access rules. Research institutions and organizations are more likely to adopt such a system if it offers these features. Otherwise, fragmenting scientific workflows across different platforms, technologies, and methodologies based on data confidentiality levels leads to increased workload and operational complexities. Such a cost might not be sustainable, making it difficult to justify the advantages of transitioning to a new DE platform, even if it offers prominent features and services.

4.4 Challenges

As Section 1.3 anticipated the thesis's research questions, this chapter has identified and discussed what makes it challenging to address the research question related to what is necessary to consider during the design of a DE for improving scientific discovery. After introducing a classification of the scientific data with their properties and the requirements arising from the stakeholders and the domain circumstances, this section summarizes three main challenges whose solutions will be presented in the following chapters. This summary aims to generalize the challenges learned from the chemical kinetics domain to be applicable to other scientific fields.

- C1: What services and functionalities should a Data Ecosystem implement when applied to scientific data? (Section 5.3)
- C2: What are the peculiar design choices in the architecture and in the workflow of a Data Ecosystem for scientific data? (Section 5.1)
- C3: What are the prerequisites to facilitate adopting and retaining user engagement in a scientific Data Ecosystem? (Section 5.2)

The combination of the scientific data's peculiarity and requirements makes the adoption of DE not easy in such scientific domains. Therefore,

Property	*C1*	*C2*	*C3*
Low-Volume High-Cost	X	X	X
Uncertainty	X		X
Accuracy Consistency	X		X
Heterogeneity	X	X	
Completeness	X		X
Reproducibility Transparency		X	X
Requirement	*C1*	*C2*	*C3*
User Engagement	X	X	X
Costs		X	
Confidentiality	X	X	X

Table 4.4: *Mapping the involvement of the properties of the scientific data and the domain requirements onto the three challenges.*

there is no unique intervention area to address this problem, but it is a multi-faced problem. Table 4.4 summarizes the involvement of a property or requisite of a scientific domain within a challenge.

CHAPTER 5

Data Ecosystem architectures for scientific data

This chapter addresses the second research question presented in Section 1.3 after Section 4.4 has formulated the corresponding challenges as a result of the analysis of scientific data properties and domain requirements. While a single organization may support the upfront costs of initial development and data collection, maintaining a data-centric system incurs a number of operational expenses to guarantee the availability and adequate performance of the system. These costs involve not only maintaining the server infrastructure to provide services and data but also managing the data itself. If the price of storing data and providing services is too high, an option is the creation of an ad-hoc organization and requesting a fee from its member to sustain these expenses. However, this approach is bureaucratically challenging and discourages the use and interest in adopting the Data Ecosystem (DE). Another alternative solution is a federated or distributed DE. In these settings, each group interested in the DE services or data can create a copy of the repository or the system, eventually share data, manage them as preferred, and dedicate as many resources as desired. It is highly scalable, but it is particularly challenging to maintain all the databases synchronized

and the software updated and ensuring a continuous willingness to share data among the participants.

The proposed solution consists of a DE with central management but a federated infrastructure. Therefore, following the categorization of DE presented in Chapter 2, this solution is a hybrid configuration between an organizational and federated DE. The proposed solutions carefully balance the design principles of a general DE with the requirements presented in Section 4.3. In the remainder of this chapter, first Section 5.1 presents the mitigations to the challenges in terms of data management and governance aspects. Data governance is a broad concept, but it is mainly related to the policies established to guarantee that data is available, accurate, secure, private, and usable. It specifies the actions, processes, and technologies that people must embrace throughout the data life cycle[1]. Data governance is a multidimensional term used on both a macro and a micro level. The macro level is related to the political and organizational aspects, while the micro level concerns the data management aspects. This chapter discusses both levels but with particular attention to the technical (micro) aspects. Since the challenges are interconnected, the data management solutions are likewise interrelated. Therefore, each proposed solution indirectly influences the resolution of other research questions. It follows in Section 5.2 a brief introduction to the architecture of such DE. The chapter concludes in Section 5.3 with the presentation of Scientific Experiments and Models (SciExpeM), the DE developed as a case study.

5.1 Management and Governance

Centralized data management within a DE helps strengthen user engagement by providing a single, reliable source of data and services. Developing a DE with centralized management but a federated infrastructure requires implementing some data governance policies in different data management aspects. This section introduces a few of them, while the remaining are then discussed in the next Chapter 6.

5.1.1 Sharing and FAIRness

Data are the key element in a DE, and its sharing functionalities help keep the system active within a community. In the case of scientific data, their value and scarcity make sharing even more crucial. Findable, Accessible, Interoperable, and Reusable (FAIR) [187] data principles have shown to

[1] https://cloud.google.com/learn/what-is-data-governance

bring many benefits to DE. Following the recommendation from the literature [188], this section presents appropriate functionalities for each FAIR principle implemented for the experimental data inside the DE.

Findable

Experiments are stored and used inside the DE through a relational database that is flexible and easy to maintain. Nevertheless, a database representation of the experiments is not findable. For this reason, for each experiment, we create an XML representation of the experiment following an Extensible Markup Language (XML) schema that is widely accepted in the scientific community of the experiment's domain. The file is then automatically uploaded to Zenodo[2] to assign to it a Digital Object Identifier (DOI) together with other metadata that make the experiment searchable without necessarily using our DE.

Accessible

Experiments inside the DE are identified both with a (numerical) primary key and the associated DOI. A primary numerical key makes implementing the relational instances in the database easier even before the DOI has been generated. The DE offers data management services through a Hypertext Transfer Protocol (HTTP) Application Programming Interface (API), accepting typical formats of the request such as Comma-separated Values (CSV), JavaScript Object Notation (JSON), and XML. One of the advantages of such HTTP API micro-services structures is that the final users are not requested to use a particular software or programming language or technical expertise to access data and services, and they can combine them as preferred. Authentication is required to use the API upon a free sign-in request procedure. Authentication enables traceability and accountability of the operations and helps keep a quality level of the scientific repository with respect to an open-access configuration.

Interoperable

Experiments in their XML representation format are a plug-and-play solution. Every researcher can use them as preferred, paying attention to the definition of each XML tag. If the experiments are accessed through the HTTP API, the same vocabulary of the XML representation format is used to query the database and for the responses.

[2] `https://zenodo.org/`

Reusable

One of the primary purposes of the DE is to reuse data, encourage their sharing among institutions and avoid duplicates. Experimental data can be uniquely cataloged through some metadata. Developing the database around the uniqueness constraint of these metadata allows to maximize the reuse.

5.1.2 Authentication, Permissions and Roles

DEs must pursue the open science policy, but authentication is fundamental to prevent malicious uses. At the same time, authentication is critical to log and trace the event in a DE. Authentication also ensures the identification of users, thus allowing to define privileges and roles and making available the designed resources from the proper organization.

The DE implements a set of policies for inserting, editing, controlling, collecting, and deleting data and using the DE services. These correspond to user privileges defined by a *supervisor* of each organization. The recommended policies follow and concur, increasing and retaining the number of users and new data. By default, each user can insert and collect data or use services. A user with editing and deletion privileges can only delete data inserted by its organization.

The open science policy coexists with confidentiality, which will be discussed in detail in the next section. The platform services are available to the entire DE community, but they consume organization resources to run. Therefore, an organization supervisor can grant service access to specific users based on their credentials.

Authentication can be used to implement a reward or token policy to incentivize good contributing behavior of the system. A user can gain tokens for its organization appropriately verifying and providing new data and consuming them collecting data. The token counter is at the organization level. Bad sharing behavior of an organization can be limited by limiting the data collection and services used for a certain time if the ratio between the consumed and produced data is lower than a threshold.

Several organizations collaborate within a DE. An *organization* is an abstract concept that groups several users. Sometimes it is possible to map this concept to other familiar entities such as a university, a research center, a department, or a research group. Each user belongs to at least one organization to be part of a DE and has at least one role.

The DE has the role of *publisher*: as soon as a content item is made open content, the DE generates, in the case of experiments, an XML representa-

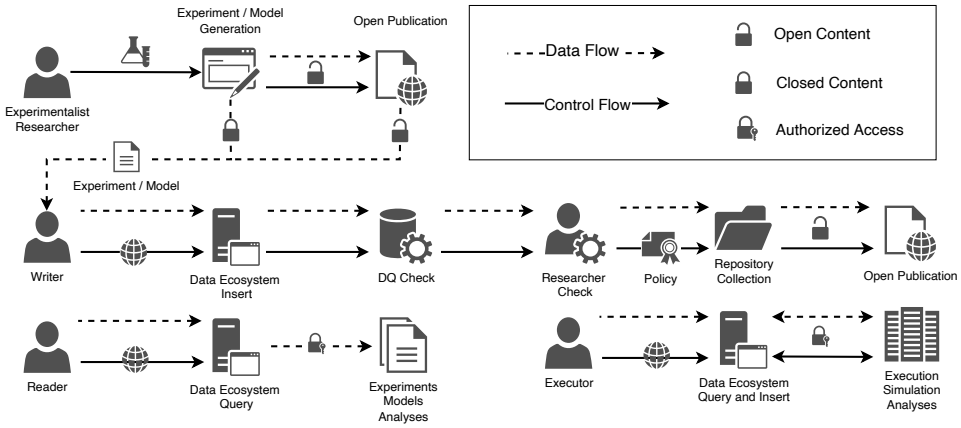


Figure 5.1: *The main five roles inside our DE, and their main four activities.*

tion file that is published in Zenodo[3] to associate a DOI to it and enhance accessibility and findability.

Besides the publisher role, for our scenario, this thesis has identified five user roles as follows. Figure 5.1 shows the five roles involved in four typical actions in the overall workflow for the model development process. The actions represented are the experiments or models generation and insertion into the DE; the collection of data, such as analyses, experiments, simulations, and models, together with the creation of simulation and analysis jobs.

Experimentalist

This role identifies a scientist who carries out the experiment and generates the experimental data. Based on the situation, the experimentalist can decide to immediately publish the results in a journal (or similar) or provide the data directly to other entities through private communication and publish them later. Accordingly to this choice, the experiments have an open or closed content policy, respectively. Even if a journal is not open access or requires a subscription, its experiments are considered open content because they are publicly available material.

Researcher

The researcher has mainly two functionalities in our DE and scenario. First, it generates the predictive model, and, as in the case of the experimentalist, it has the faculty to choose the publication policy. Second, it has the duty

[3] https://zenodo.org/

to verify the experiments in their validation procedure as described before. Suppose the experiment that has to be validated is open-content. In that case, a cross-validation strategy is preferred: a researcher from a different organization of the experiment ownership will perform the task to avoid possible bias and enhance the DE's overall trustworthiness. It is assumed that there is at least one researcher per organization.

Reader

The reader represents the user that has permission to access the open contents and all the closed contents belonging to its organization. Thanks to authentication, transparently, it is possible to hide part of experiments, models, simulations, and analyses without changing the API.

Writer

The writer is a trained user who has the task of inserting all the collected data into the DE. It is a trained user because, for this field, it is not a straightforward operation, and it requires basic domain knowledge, even if the system and the researcher will check their validity later. The writers mainly insert experiments and models. They can find these data in the literature, or they can be provided through private communication. In any case, they are responsible for associating the correct content policy with objects.

Executor

This role represents a kind of user that has the privilege to allocate resources and generate new data in terms of simulations and analyses. In both cases, the executor needs to have access to both experiments and models to create a new simulation or perform analyses (like in the case when it is needed to compare experiments against simulations). This kind of operation could result in expensive operations. Also, in this case, domain experience is required, for example, to set the optimal numerical configuration to solve a simulation numerically and thus use the computational and storage resources wisely. It is worth mentioning that even if an experiment is closed-content and the user does not have the permissions, its metadata, i.e., in this domain, the experimental condition, is, in any case, open, and therefore it is possible to simulate this configuration. Nevertheless, all the analysis operations concerning comparing the simulated data against the experimental data will be hidden.

5.1.3 Confidentiality

A trustworthy DE is one of the requirements for achieving high system engagement. From the security perspective, it means ensuring authentication and confidentiality. Authentication is more related to the business process organization, user roles, and privileges. On the other hand, data confidentiality is related to the management of user interactions across different organizations within the DE in the different stages of the business process. Data confidentiality addresses the domain requirement for scientific data that need to use the DE services and ultimately share the data, but, on the other hand, for some time, during model development, it also needs not to be publicly accessible to guarantee academic and industrial advantages. Requiring all data to be accessible to any user of the DE at all times could potentially disrupt the workflow of a research team based on the usual scientific data management policies. Poor data management policies could ultimately discourage the adoption of the DE. In the scientific data domain, experiments and models are the main types of data that require confidentiality. In particular, new experiments are confidential for the reported values, while their metadata do not have confidentiality constraints. The metadata of the scientific data should be accessible without any restriction. For instance, the metadata of an experiment describes the experimental setting. Knowing all the experimental settings discourages other researchers from investing in performing an experiment that others have already investigated, even if it has yet to be made publicly available, optimizing the resources and reducing duplicates. Within organizations, groups of users that work together and thus share the same resources can be defined. Examples of organizations are universities, research groups, departments, or research centers. Every user in the DE is affiliated with at least one organization. Some examples of data confidentiality policies are presented below. All the scientific data published or generated by a user belong to his/her organization(s). The user, during the data publication, has to specify whether the data are open or closed and under which conditions. All the closed-content data will become open, for example, after an embargo of one year from its insertion in the DE. Within the DE, each user could access all open-content data of all organizations and all the closed-content data belonging to his/her organization(s). In terms of implementation, data confidentiality can be ensured by encrypting the confidential information with the public key of the owner of the confidential data. The open science policy coexists with confidentiality. The accessibility of confidential data can be granted to other users belonging to a different organization of the confidential data

ownership. In fact, the configuration in organizations described previously allows an easy share of closed-content resources among them with different levels of granularity and relationship: a single experiment or a group of them could be shared with another organization, or an organization can share in one direction or both directions the whole closed-content data. Another solution to address the problem of confidential data would be to use multiple databases, one for the open data and as many other databases as the number of organizations participating in the DE. This database could stay physically in the organization itself but increase the sparsity of the data and disincentives the creation of a data collaborative. Moreover, it is technologically challenging to keep synchronized multiple databases.

5.2 Architecture

A DE for scientific data requires specific design choices regarding the overall architecture illustrated in Figure 5.2. First, it is fundamental to understand the business expectations when employing a DE in a scientific domain. It enables identifying the relevant scientific data and characteristics, required services, and actors involved in the business process. The unique characteristics of the data guide the definition of the database schema. As previously discussed in this chapter, data management should be centralized. This architectural decision aims to enhance trust, transparency, traceability, and efficiency. The technological implementation of the database could be distributed, even if it is not recommended, due to the risk of consistency issues. Furthermore, central data management encourages users to share data, promoting collaboration and knowledge exchange within the DE. Figure 5.2 depicts four types of entities: *data ecosystem*, *user*, *third-party service*, and *worker*. They communicate with each other through the internet, in particular with the HTTPS protocol. The DE offers its services to the users through microservices through API endpoints. This service-oriented architecture allows flexibility, extensibility, and high maintainability, and users can request and combine services as preferred. All the services of the DE are available through authentication provided by the back end. Such authentication prevents malicious usage of the system and allows giving users different privileges and permissions. Through the back end, the DE can offer services with a combination of legacy and developed ad-hoc modules, both on-premises or in-cloud, with *third-party service*s. Finally, to maintain the scalability (please refer to Section 5.3.4 for more details) of the system, the architecture foresees delegating the computational burden to external *workers* where the DE coordinates and distributes

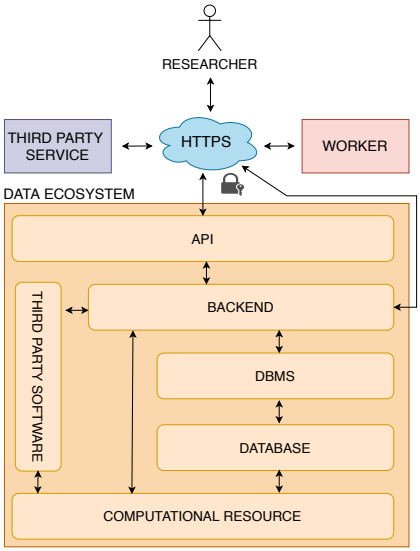


Figure 5.2: *General overview of the architecture and the main actors of a DE for scientific data.*

the workload. From this point of view, the infrastructure is provided as a DE federated configuration.

5.3 SciExpeM

This chapter presents SciExpeM [4]. It is the DE developed for the case study of this thesis in chemical engineering starting from a prototype, in which the primary services, the functional requirements, and the architecture to support the needs emerging from the large-scale data-driven validation of scientific models were discussed [189]. More specifically, it is a DE to support and improve the predictive model development process of chemical kinetics models. It follows the architecture in Section 5.3.1, and in Section 5.3.2, a detailed description of the process that SciExpeM improves and supports. Then, Section 5.3.3 presents the implemented services, while Section 5.3.4 explains how SciExpeM implements a scalable DE to meet the sustainability requirements presented in the previous chapter.

5.3.1 Architecture

NIST Big Data Public Working Group presents the NIST Big Data Reference Architecture (NBDRA) guide that describes, using a functional com-

[4] `https://sciexpem.polimi.it`

ponent view, the roles with their actions and the components that carry out the activities for a Big Data architecture [190]. According to these guidelines, this section presents in more detail the SciExpeM architecture, also shown in Figure 5.3.

Figure 5.3 depicts a *System Orchestrator* that coordinates the configuration and management of the other components of the Big Data architecture. In our scenario, the *System Orchestrator* corresponds to the management of SciExpeM.

SciExpeM, instead, is *Big Data Application Provider* that encodes the business logic and executes a specific set of operations to the data. SciExpeM implements functionalities for collecting, preparing, visualization, and access control to the repository and other services. According to the following Section 5.3.2, the main functionalities that are included in the SciExpeM architecture in Figure 5.3 are *analysis*, *validation*, *control*, *insert*, and *simulation*.

The union of four submodules represents the *Big Data Framework Provider*, with an additional transversal layer representing the HTTPS REST API communication mode with the SciExpeM framework and a module of resource management and optimization. The implemented REST API follows the recommendation of Rodriguez et al. [191]. Beginning from the bottom to get to the top, Figure 5.3 shows a layered structure of the DE. The system functionalities are offered through the *Services* that are supported by *Database Manager* that utilizes the *database*. All of them need the *Computing Resources* to run.

Finally, the *Experiment Source* represents the system's *Data Provider* in terms of both literature, private communications between research labs, or experiments and predictive models entered by users. Instead, the *Researcher* is the *Data Consumer*, representing all the typologies of the user that interact with the system to manage it or request services. The systems' interactions occur through a user-friendly interface or using the API.

The NBDRA defines three types of arrows. The *Data* arrow represents the data flow between system entities. The arrow *Software (SW)* represents the transfer of software tools for data processing. For example, it is necessary to provide software tools to interpret the data provider's experiments in different formats. Finally, the *Service Use* arc indicates all the programmable interfaces between the system's various entities.

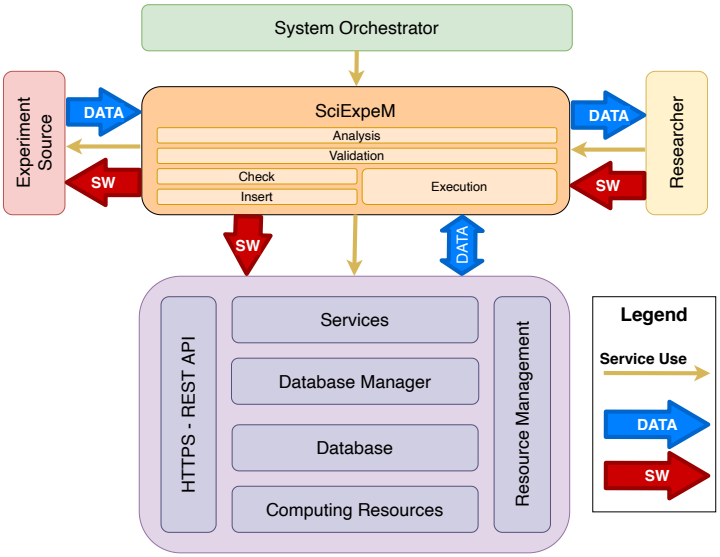


Figure 5.3: *Sketch of the architecture adapted from the NIST Big Data Reference Architecture (cfr. [190]) for the management of experiments to support the development of predictive models.*

5.3.2 Chemical Kinetics Model Development Process

Chapter 3 has introduced a general model development process in a scientific domain from a business perspective. The adoption of a DE to support such a process has as a benefit a general speed-up of the overall process and its refinement. Thus, the DE aims to deliver, in a faster way, better predictive models. The process is improved both in terms of reliability, for instance, automating human-error-prone tasks, and trustworthiness, introducing new aspects in the process. Therefore, this is an example where the information system improves the pre-existing business workflow after analyzing its characteristics.

As explained in Chapter 3, the model development process is a data-driven procedure. Therefore, even if the final scope is to generate predictive models, this process has two protagonists: experiments and predictive models. Both of them are fundamental to starting the model development process and undergo a similar process to be included in the repository of SciExpeM. This procedure comprises four stages as shown in Figure 5.4 and described in the following.

1. *Collection.* A researcher can have the necessity to include a new experiment or predictive model for different necessities. For instance,

he/she wants to verify if a model can correctly represent new experimental results or is reliable enough to be used in real-life applications, or needs improvements. Experiments and models are collected following two possible events: private communication and literature review. In the former, the data is not yet available in the literature and need to be treated with confidentiality. In the latter, the researcher has found new data that can be included in the repository as a result of a literature review. The data can be found in many formats, both machine and human-readable. These formats include CSV, JSON, XML, structured or unstructured xls, or general text files such as pdf or doc.

2. *Insertion*. Both predictive models and experiments, once collected, are inserted in SciExpeM through a dedicated service. Based on the ontology of the inserted data, SciExpeM requires all the mandatory information. The completion of some fields is automated. For instance, the literature reference of an experiment. Thus, the collected data in various typologies and formats are translated into machine-interpretable forms.

3. *Control*. It is counterproductive to validate a predictive model against a wrong experiment and vice-versa. For this reason, SciExpeM controls all the data inserted in the repository. This stage was not included in the original business process of model development. Instead, it is a critical stage since it enhances the quality of the repository and the trustworthiness of the sharing platform. As the first step, the system checks if the data is already included in the repository. Then, based on the pre-defined data quality rules, performs automatic controls. If the quality controls are not successful, the data is rejected. Otherwise, the data is stored in the repository with an *unverified* status.

4. *Verification*. The syntactic and semantic checks in the previous step can identify gross errors, but due to the domain complexity, it is not straightforward to control all the aspects. Thus, since the quantity of the data is limited, a manual and visual quality check performed by an expert can ulteriorly improve the quality of the repository. If the data succeeds in the verification step, its status is changed accordingly (textitverified), otherwise is rejected.

Each activity requires different levels of knowledge of the domain, so different roles can be identified within the process since there is no exact mapping between the qualification of users (student, intern, researcher, etc.) and the role responsible for a task. Duties are assigned every time, and for

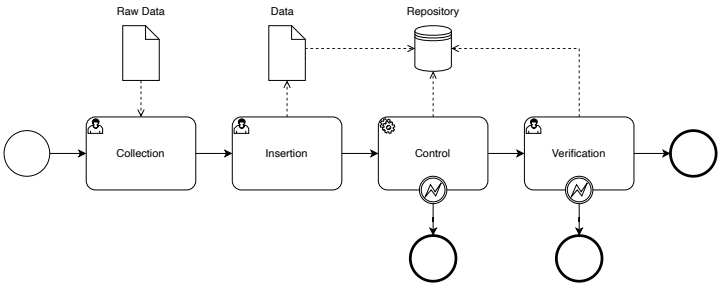


Figure 5.4: *Process of the experiments and predictive models to be included in the SciExpeM's repository.*

this reason, the need emerges to provide great flexibility in defining the roles and permissions in the system which is supporting the process.

Figure 5.5 presents the *predictive model development loop* within SciExpeM. It foresees five main phases as follows.

1. *Simulate*. Given a collection of experiments and a model, the first step of the loop is to simulate the initial conditions expressed by the experiment collection with the given model. If the simulation has not been computed yet, the simulation is actually performed. Then, the results of the simulations are stored within SciExpeM to enhance the reuse of data and save the computational cost of running a simulation. It is necessary to prepare specific input configuration files for each experiment in a format comprehensible by the simulator, specifying all the characteristics necessary to simulate the experiment.

2. *Validate*. This step involves comparing the results of the simulations with the experimental data. This procedure was typically conducted in *qualitative* way: an expert compares, based on his/her experience, the simulation results against the experiments. Different experts can have different opinions on the comparison. SciExpeM automates this time-consuming and ambiguous procedure using a *quantitative* approach. SciExpeM adopts Curve Matching as a quantitative tool that can measure the function shape similarity between the experimental and the simulated data. The result of the validation is a kind of analysis results (Section 4.1) that are stored within SciExpeM and can be further elaborated in the next stages. During the validation, some outliers may appear: if a simulation diverges significantly from the experimental data, the experiment is very uncertain, or the model needs improvement. Since disambiguating this operation requires expertise,

this operation is manual. If an experiment is unreliable, the status of the experiment will be changed to *invalid* to exclude it from further utilization.

3. *Evaluate*. Based on the predictive performance of a model, the model developer decides whether he/she is satisfied.

4. *Analyze*. If the model developer is not satisfied, now SciExpeM introduces a new analysis stage. In the previous business process, it was completely skipped. Instead, SciExpeM, with a collection data analysis service, can automatically extract systematic information from the analysis result data and guide the next model improvement.

5. *Generation*. Once the analysis results are available, it is possible to improve the model to cover the gap from the experimental data. The model for chemical kinetics is hierarchical and modular. This feature simplifies the simulation and allows researchers to work independently on more modules at the same time. Each module covers a portion of the domain regarding a specific species with precise time scales and quantities. The following steps are:

   (a) *Module Selection*. Various reasons can bring to the decision of which modules of the model to improve, including social or industrial contexts or the availability of new experiments on a given fuel are the main reasons.

   (b) *Theoretical Study*. The model's theoretical development begins, trying to understand which reactions occur, estimating or calculating the constants of the reactions depending on the problem's complexity.

   (c) *Integration*. The developed module is translated into the simulator model format and integrated into the combustion kinetics model.

   This thesis, with Chapter 7, discusses how to automate and improve the *generation* stage using data science methodologies.

5.3.3 Services

SciExpeM adopts a modular service structure that is easy to extend. New services can be offered through new endpoints in the API, implementing new functionalities, or combining the pre-existing ones. Figure 5.6 shows the dependencies among the current functionalities represented by an arrow,

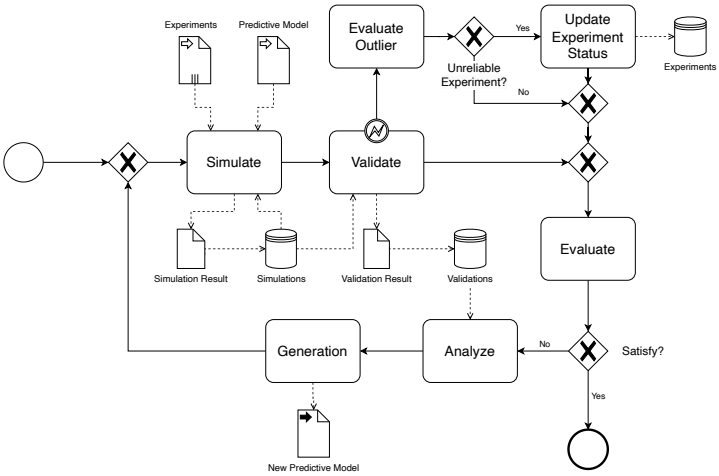


Figure 5.5: *Predictive model development loop within SciExpeM.*

pursuing the decoupling and reuse principle of a microservices structure [192,193]. A characteristic of SciExpeM derives directly from this strategy: providing essential services to the end-users, then they can combine them as preferred.

In the following, each paragraph describes a group of services, in particular:

1. *Manage Scientific Data* This group of services takes care of the management of scientific data.
2. *Control*: This set of services tries to keep high quality in the SciExpeM scientific repository.
3. *Development*: This group of services offers functionalities to investigate and improve a predictive model.
4. *Other Services*: This collection of services represents all the other services implemented by SciExpeM.
5. *External Services*: These services are external functionalities or software that are not directly implemented in SciExpeM.

Ensuring convenient accessibility to SciExpeM services remains a crucial goal. While employing communication endpoints for interface purposes offers great flexibility, this approach might not be convenient for all users. Two alternative methods of engaging with the system have been conceived to address this concern, both leveraging the existing endpoints.

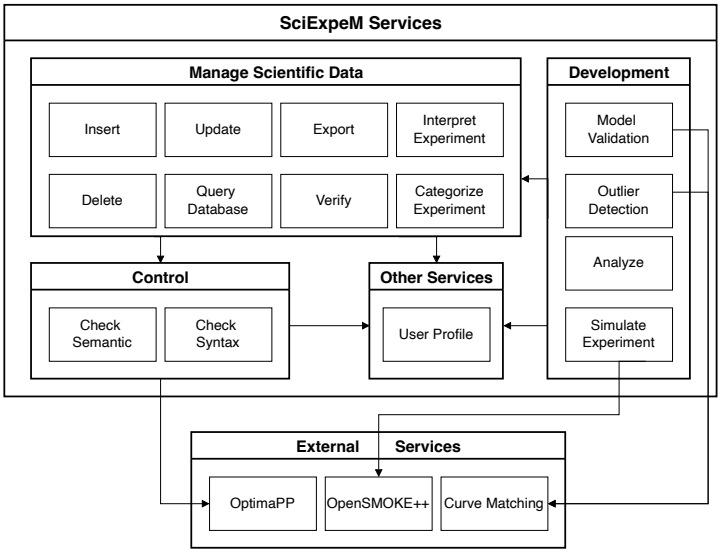


Figure 5.6: *Main services of SciExpeM system with their dependencies.*

Firstly, users can interact with the system via a web interface, facilitating tasks such as the insertion of new experiments or predictive models. Alternatively, users can leverage a Python library that wraps the database tables illustrated in Figure 4.2 into Python objects and the services into function calls without explicitly interacting with the HTTP API.

Manage Scientific Data

This set of services represents the system's functionalities to support the life cycle of scientific data within the SciExpeM repository.

SciExpeM accepts new experiments and predictive models through the *Insert* service. Scientific data can be collected in many machine-readable representation formats, and SciExpeM accepts the most common ones. However, SciExpeM stores the scientific data in a relational database whose schema follows a pre-defined ontology. To accept data in other formats, SciExpeM implements translation engines. At the same time, this characteristic allows for easy export of the data from the system to other environments through the *Export* service in the desired format.

Once scientific data is in the repository, other services are critical for maintenance and consultation. *Update Experiment*, *Delete Experiment* represent the services to update or delete scientific data from the system. However, to use these services, particular attention is needed. Updating, for

instance, an experiment is intended to rectify errors entered during the insertion of data and not modify the experiment itself. Since a DOI is associated with an experiment, any modification from the original state would invalidate the correspondence with the DOI. Any writing interaction with databases triggers the *Control* service.

Query Database represents the service to query the repository to retrieve scientific data.

As discussed previuosly in Section 5.3.2, automatic checks provided by *Control* services reduce the possibility of errors, but they cannot be completely detected due to the domain's complexity. In this case, an expert is required to *Verify* the new inserted data, changing the status accordingly.

A verified experiment can be then used by a variety of others services. However, SciExpeM must interpret correctly the semantic of an experiment to know, for instance, which and how to compare the experimental to the simulated data. This seemingly logical and straightforward problem, but it is not easy to automate in a scientific domain. The *Interpret Experiment* service address this aspect.

As explained in Section 4.1, experiments are records of measured properties and other metadata like the instruments used, the authors, etc. Besides, among the measured properties is not rare to find also additional measured information that specifies, for example, the environmental conditions of the measurement. However, there is no clear way to distinguish which measurement is the subject of the experiment. Thus, the semantic heterogeneity of the measured data generates ambiguity and makes it impossible to automatically distinguish the experiment's actual subject (primary data) from the other measurements (secondary data) without additional domain knowledge.

For this reason, it is necessary to define a flexible methodology to distinguish the primary information from the secondary data in a complex database model. In other words, it is critical to define an approach to transfer the domain knowledge into the SciExpeM to interpret the semantics of an experiment correctly and treat all the database entries with equal semantics in the same way.

To explain this problem, it follows a simplified version of one scenario coming from our case study. For each experimental data there are always two measured quantities: pressure and temperature. Our database model (NoSQL) for experimental data has, for this reason, two data entries (PostgreSQL array fields) that store the temperature and the pressure, but it is impossible to know the primary data between the two. In fact, sometimes the primary data is temperature, sometimes it is the pressure. In other words,

it is fundamental to distinguish the independent, the dependent and the accessory measured variable, when there is no fixed relationship. Fortunately, this ambiguity can be overcome by checking a variable number of experiment metadata. Manual management of this complex database is not feasible because an experiment could contain dozens of measured properties, and, for example, we should tag each of them correctly if they are primary data or not. Moreover, this procedure should be repeated hundreds of times, once for each experiment, making it hard to analyze a large amount of data.

To address this problem, this thesis propes a dynamic and automatic interpretation of a database model based on rules, similar to what is done for data cleaning or to ensure consistency and accuracy in a database [194]. Given a model $\mathcal{E}$, that is an abstract representation of a model affected by ambiguity, we have to assign, for each entry $e \in \mathcal{E}$, an *interpreter* entry of the model $\mathcal{I}$. This model can save additional meta-information that could be useful for other tasks. Each interpreter knows how to distinguish the primary data from the secondary information and correctly map them. This is possible because the interpreter has multiple references $M = \{m_1, ..., m_n\}$ to a *mapping* model $\mathcal{M}$ that knows, for example, the correct relation of dependent-independent variable, or more in general, can separate the useful information from the secondary one, and if necessary, pair them. In order to associate an interpreter to an entry of the model $\mathcal{E}$, we have to associate a set of rules, $R = \{r_1, ..., r_k\}$, to an interpreter. These rules r are entries of another table in the database, *rule*, $\mathcal{R}$, where each element specifies a name of the model N, the attribute's name A and value V. A rule $r \in \mathcal{R}$ is fulfilled by an entry $e \in \mathcal{E}$ if A is an attribute for e and the corresponding value of the attribute is equal to V. The model name N is an optional field that, if defined, specifies that the rule is not directly on an attribute of the model e, but it is related to an attribute of another model N that has a reference to the entry e.

If an entry e fulfill all the rules r associated to an interpreter $i \in \mathcal{I}$, we can associate the interpreter i to the entry e.

Figure 5.7 shows a toy example. In this case, the table affected by ambiguity is the *Data* table, storing data related to an experiment in *Experiment (Exp.)*. In particular, in this case, it is important to distinguish which is the independent variable. Each entry of the *Exp.* model has a reference, *INTER. ID*, to an entry of the model *Interpreter (INTER.)*. To assign the correct interpreter to each entry, all rules from the model *Rule* related to an interpreter through *INTER. ID* field should be respected by the entry of *Exp.* model. A rule is fulfilled by an entry of the model *Exp.* if, in correspondence with the attribute specified by the rule in terms of attribute

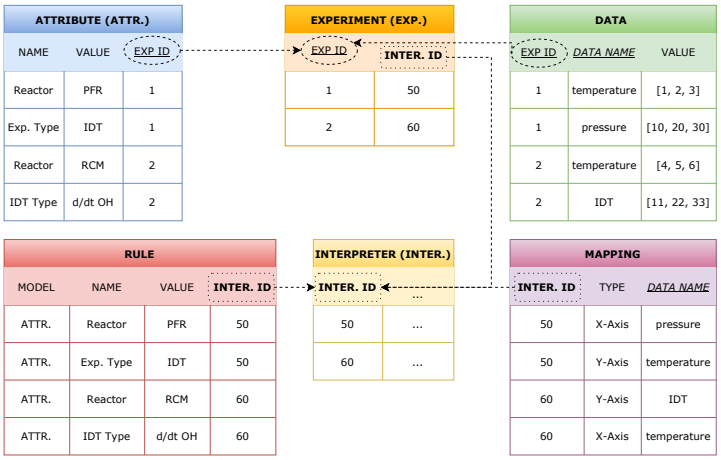


Figure 5.7: *An example of the rule-based interpretation. INTER. with ID 50 is assigned to the Exp. with ID 1 because it fulfill all the rules (Rule table) associated with this interpreter. For instance, the Exp. with ID 1 has an entry in the Attr. table where the content of the name and value fields are respectively 'Reactor' and 'PFR' as requested by the fields of the rule. The interpreter tells us the role of the entries in the Data table related to an experiment. In this case, the pressure is the x-axis, and the temperature is the y-axis as specified by the Mapping entries related to Inter. with ID 50.*

name and value, the entry has the correct combination of *Attribute (ATTR.)*. Once the interpreter is assigned, the reference to the *Mapping* model helps distinguish, for example, the primary data or the independent variable.

Control

Every data in the repository has to be syntactically and semantically correct. SciExpeM performs some controls and refuses the scientific data if the controls are unsuccessful, notifying the user. The platform executes the control on the data every time a service interacts with the database in writing mode. The controls happen automatically and transparently without invoking the service explicitly. SciExpeM adopts OptimaPP to check the syntax of the experiments [195]. This software controls whether the rules to define a chemical kinetics experiment in the ontology defined by the ReSpecTh format are fulfilled [107]. To assess semantic errors in the data, SciExpeM can automatically check some simple essential characteristics. For instance, the agreement between the unit of measurement and the measured property.

Development

The analysis is a central task for improving the development process of scientific models. As previously explained in Section 5.3.2, the first step in this macro phase is to *Simulate Experiment*s with a predictive model. The second step includes comparing the simulation results with the experimental data to measure the predictive model's performance using the *Validate Model* service. This service quantitatively computes the predictive model performance. SciExpeM adopts *Curve Matching* to measure the difference between the predictive model's simulation results and the experimental data. The validation results are then used by the *Analyze* service to extract information using data science techniques to guide the next predictive model improvement. *Outlier Detection* is an example of an *Analyze* service. It provides more precise information regarding the model's anomalous behavior in some portion of the domain, collaborating with *Categorize Experiment* service. SciExpeM automatically performs the analyses with the *Interpret Experiment* service that provides the necessary knowledge to correctly compare the simulator's results with the experimental data.

Other Services

SciExpeM relies on other services that transparently provide additional functionalities to support, for instance, the user interface, other services, or the control over the platform itself. The *User profile* is a critical aspect of the system since it allows users to request a SciExpeM service. SciExpeM uses authentication and service permission management to guarantee a secure working environment, reliable scientific data, and correct use of resources. For each service, the policy is to define its corresponding permission. The permission can be organized in groups to facilitate their management. Therefore, when SciExpeM offers a new service, it is necessary to associate it with permission and then add the corresponding authorization to users needing access. Finally, among the *Other Services*, a logger functionality keeps track of all events in the DE.

External Services

SciExpeM adopts external software or service to provide its functionalities. It executes these services and collects the result from them as a black box. For instance, OpenSMOKE++ is used as numerical simulator for the experiments [182] , OptimaPP to check the experiments in the ReSpecTh format [195], and Curve Matching (CM) to measure the similarity between two sets of data [121].

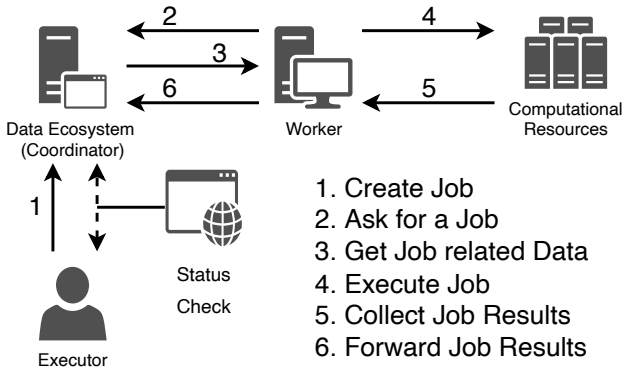


Figure 5.8: *Coordinator-worker architecture.*

5.3.4 Scalability

A predictive model can theoretically simulate an infinite number of domain conditions. Similarly, using the analysis tools and combining them as preferred, it is likely to generate a vast number of analysis data. Neglecting the space needed to store such quantities of data, the first limitation that makes this idea unfeasible is the amount of computational resources needed to generate them. A centralized DE where all the computational burden is on a single entity is not sustainable. Even if the cost is shared, the bureaucracy behind sharing computational resources is very complicated. Moreover, as previously discussed in Section 4.3, this practice can discourage the adoption of the DE. A possible solution is a coordinator-worker paradigm where the DE, i.e., the coordinator, collects the jobs and distributes them among the workers, that in some cases can delegate the job to other machines as shown in Figure 5.8. The coordinator-worker configuration is scalable and allows each organization participating in the DE to decide how many computational resources to dedicate and use only for their jobs.

In addition, simulating an experiment can take anywhere from a few seconds to several days. For this reason, it is crucial to fulfill two requirements. First, the simulation of an experiment must not be a blocking request. Second, it is also necessary to verify that a simulation has not already occurred or started to guarantee data reuse and save resources. Figure 5.9 shows a Business Process Model and Notation (BPMN) that describes the interaction between various services and entities to perform a simulation fulfilling the previous requirements. When a user submits a request to start a simulation, the system creates a transaction. Within the atomic transaction, SciExpeM checks that there are no other simulations already completed or

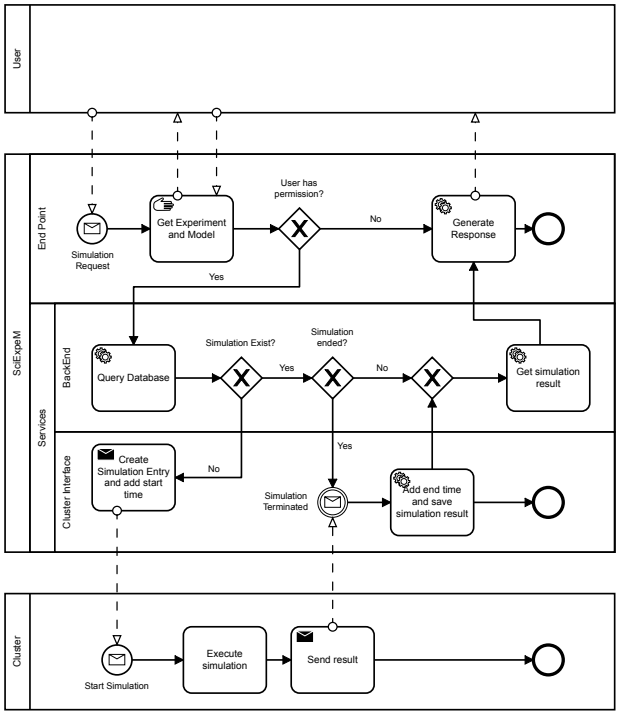


Figure 5.9: *BPMN for the request of an experiment simulation.*

started. In any case, the system replies with the result, if available, otherwise with a response containing information regarding the simulation's status. Referring also to Figure 4.2, if the request regards a new simulation, the system forward the simulation to the worker after updating the corresponding *starting time* field of the database's entry. When the worker finishes the simulation's execution, the system saves the results and updates the simulation *ending time*.

CHAPTER 6

Data Preparation

Data preparation is one of the most important aspects of governance and management of a scientific repository. As explained in Chapter 3 and in Section 5.3.2, the four types of scientific data presented in Section 4.1 are tightly connected in the predictive model development process. The purpose of data preparation is threefold. First, since the predictive model development procedure is a data-driven methodology, the quality of data discussed in Section 6.1 can significantly impact the final results. Section 6.2 introduces how the uncertainty of the scientific data is related to the data quality dimensions, and it introduces a method to predict it. Second, since the model development process is the result of a complex procedure, it is important to provide all the necessary information to the researcher about how this model has been developed. Therefore, Section 6.3 discusses how to make this process transparent and thus enhance the trust in Data Ecosystem (DE), using data provenance. Finally, the third aspect presented in Section 6.4 discusses how and why it is important to assess the diversity of the dataset for the validation of a predictive model.

6.1 Data Quality

Data have a central role in all data-driven applications, and their quality is critical since it directly influences the reliability of all the downstream uses [196]. If the Data Quality (DQ) rules are correctly set, they mitigate the typical Garbage In - Garbage Out (GIGO) hazard of all data-driven applications and the fast spread of wrong information in processes where data are linked [197]. According to the process described in Section 5.3.2, the four types of scientific data, experiments, predictive models, simulation, and analysis results, are strictly related. Therefore, the scientific repository must ensure a certain DQ level. Without proper DQ control, unreliable information will spread rapidly and negatively affect the other data types in the DE. In the last decades, research on DQ has defined analysis dimensions and metrics to define and assess the quality of data. DQ identifies dataset characteristics and presents quantitative measures of the corresponding quality dimensions. In the end, DQ quantifies and highlights the strengths and criticalities of a dataset. Over time, hundreds of different DQ dimensions were defined, each quantifying a different quality aspect of the data [9].

Nowadays, predictive models are increasingly data-driven, even in domains where a description with physical laws of the phenomena is available. For this reason, DQ plays a more and more central role in the model development process since it directly impacts the prediction quality. In addition, as previously discussed in Section 4.3, ensuring certain DQ levels within the DE enhances the platform's trustworthiness, thus starting a loop of increasing the number of users as a consequence of the increased amount of collected data and vice versa. Following the *fitness for use* concept [9], a DE for the development of data-driven models based on experimental data needs to consider completeness, consistency, and accuracy as DQ dimensions since they are the most widely used across different domains and provide a good assessment of the quality of data products. Timeliness is not of interest in this thesis scenario, even if it is often used as a quality metric, mainly for three reasons: first, even if older experiments are carried out with older and less precise instruments, they still represent a valuable source of information, and their imprecision should be included in their uncertainty evaluation, which it "just" needs to be handled correctly. Second, since scientific experiments are expensive and hence rare, it is pretty unlikely that multiple experiments are carried out in exactly the same conditions, thus "updating" the old values. The last one relates to the infeasibility of repeating an experiment since it is practically challenging to replicate the same

environmental conditions. For a similar reason, since the predictive models are deterministic, the simulated data does not change over time if forecast with the same model and numerical configuration of the solver.

In the SciExpeM DE, the DQ control process discussed in Section 5.3.2 is composed of two parts, one automatic and the other manual, where the automatic control is performed right after the insertion of new data in the repository and not, for example, a posteriori based on a recurrent schedule. Data that does not reach the minimum data quality requirements are immediately rejected.

Completeness The domain ontology defines which metadata that describe the data are mandatory and in which conditions. For example, according to ReSpecTh ontology for combustion kinetics [107], the unit of measurement is mandatory in every experiment. Therefore, in a DE that hosts experimental data, it is sufficient to specify a collection of rules that check the completeness of the scientific data's metadata.

Consistency By defining a list of rules, consistency quantifies whether the information stored in different parts of the database but semantically related to each other and regarding the same data is congruent [9]. For instance, considering the experiments as a scientific data type. An example of a consistency rule between the type of a measured property and the unit of measurement stored in two different database fields regarding the same experiment is the plausibility of the unit of measurement regarding the reported property. In practical terms, if the type of the measured property is "pressure", potential units of measurement are "atm", "Pa" (Pascal), "bar", etc., but not, for instance, "K" (Kelvin).

Accuracy Accuracy is a challenging data quality dimension to assess. It is related to the precision of the data in representing real-world values. Given a ground truth, accuracy measures the discrepancy between the value reported in the database and the real one. Following the previous example, a bunch of valid units is plausible for "pressure", but only one value is correct for a measured value given the unit of measurement. However, accuracy is also strictly connected to experimental uncertainty (that, unfortunately, is not always provided together with experimental data [132]). Measuring accuracy is challenging since a ground truth is needed for its evaluation. The accuracy is determined using different data sources and thresholds for numerical values. For instance, experiments do not have a ground truth because uncertainty is always present in the measurements [27], and assessing

whether the measurements are correct or inaccurate is challenging. Uncertainty gives scientific data a confidence of the reported values. In the case of experiments, the same experiment is repeated multiple times to quantify the uncertainty. However, repeating the same experiment, as said before, is challenging by itself. Excluding this problem, after an experimental campaign, the average of the measurements will be the reported value, and the standard deviation will be the uncertainty range.

Section 6.2 discusses in detail how uncertainty is related to the DQ dimensions for scientific data. In this case, the concept of accuracy is also related to the DQ dimension of consistency (or agreement) of different experiments concerning the same (or similar) experimental observation. Therefore, evaluating the consistency between different experiments regarding the same condition can be reduced to their accuracy evaluation.

6.2 Data Uncertainty

As introduced in Chapter 3, with the validation of the predictive model, the model's predictions are compared against the experimental data, and thus the predictive model performance can be estimated. This model validation step can not be appropriately performed if the experimental uncertainty is missing. Uncertainty for the experimental data can be represented with error bars, as shown with an example in Figure 6.1. Uncertainty is a discriminatory factor to establish whether the model predictions are congruent, i.e., the model predictions are inside the experimental error bars.

As discussed in Section 4.1, experiments are a particular kind of data since they record physical measurements that, by definition, are affected by experimental uncertainty, also known as experimental error [36]. This is usually obtained by repeating the experiment under the same conditions. Different sources of the same observations are hence used to build a *ground truth*, that is used to estimate the uncertainty [135]. It is more likely that only recent papers systematically report experimental uncertainties. However, this is not systematically true due to the cost of replicating them [132]. Similarly, experiments carried out in the past are more likely to be imprecise due to the adoption of more imprecise instruments to perform the measurements. However, since they are already available to the community, it is unlikely that someone is willing to invest in already present information, so, often, scientific data and, in particular, experiments are unique. In any case, with or without, with more or less uncertainty, scientific data are still valuable sources of information.

Table 6.1 is an example of a possible large uncertainty present in a

combustion experiment. The experimental data does not explicitly report the uncertainty. At almost identical temperature and pressure conditions, there is, for example, a significant variation in the measured ignition delay times. There is over a 65% difference from the value recorded at $(930\ K, 3.586\ atm)$ and $(930\ K, 3.534\ atm)$. Therefore, this is an example to demonstrate that uncertainty is always present in the experiments, but it is often not reported, although it is not negligible. Consequently, a method for predicting missing uncertainties within the repository serves a dual purpose: it enhances the overall data quality profiling and the reliability of developing predictive models.

In Chapter 2, various methodologies for predicting experimental uncertainty are presented, with a common assumption that multiple experiments conducted under similar or identical conditions are available for robust statistical analysis. However, as previously mentioned, it is important to note that the availability of multiple data points under such similar conditions is rare in numerous application domains. Consequently, these methodologies may not be applicable, necessitating reliance solely on the existing available information.

Instead, this thesis proposes a methodology that leverages only the already available information and Knowledge Graph Embedding (KGE) to estimate the missing experimental uncertainty. Specifically, it leverages three facts: first, even if they are affected by epistemic uncertainty [123], predictive models represent more or less precisely the domain; thus, they can be used to approximate the ground truth. Second, it is rare that two different experiments from different sources measure the same phenomena in the exact environmental conditions. However, experiments carried out in similar environmental conditions should report similar values. Thus, their embedding, derived from the knowledge graph, should be close. Finally, the metadata of an experiment has additional information, such as the authorship or the instruments used. It is reasonable that sometimes the aleatoric uncertainty [123] can have a systematic part due to, for example, a wrong calibration of the instruments of a specific lab. In other words, the KGE can learn hidden, systematic, and complex relationships between the metadata of the experiments and the uncertainty.

To validate and study the new proposed methodology, multiple test cases are constructed and used to train the KGE model and then assess its predictive capabilities. Finally, the procedure is applied to the real case scenario of chemical kinetics data highly affected by missing uncertainties. This contribution demonstrates that it is possible to infer the missing experimental uncertainty when enough structured information about the experiments

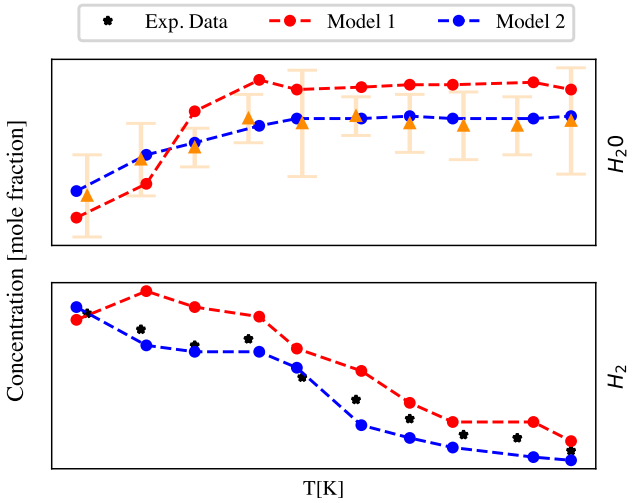


Figure 6.1: *The relationship between experimental data, simulated data, and experimental uncertainty. Experimental data without uncertainty do not allow for properly evaluating if the model predictions are correct.*

and their uncertainty is available.

6.2.1 Knowledge Graph

Given a domain ontology, a Knowledge Graph (KG) can represent a repository of experiments. An experiment is the union of information from its metadata and the reported observation. The ontology classes and properties become KG entities, and their links become the KG predicates.

Definition 6.2.1 (Knowledge Graph)**.** *A Knowledge Graph (KG) is a collection of facts, each connecting two entities with a specific relationship. Formally, A KG, $G = \{E, R, F\}$, is a sets of entities (E), relations (R) and facts (F), respectively. A fact, or triple, $(h, r, t) \in F$ is composed by th 'Head' (or subject) h , a 'Relation' (or predicate) r and a 'Tail' (or object) t. The representation format of the triples is known as Resource Description Framework (RDF) [199].*

The literature defines a metamodel to describe an experiment [107,145], and the corresponding KG is shown in Definition 6.2.1. This KG accounts for the most popular and general metadata that describe an experiment.

Not all the relevant factors for predicting experiment uncertainty may be considered in the KG. However, the model embedding validation can spot

Table 6.1: *Example of uncertainty in* `DOI:10.24388/g00000007` *experiment by [198]. In bold and underlined groups of nearby points but with significantly different measured Ignition Delay Time (IDT).*

Temperature [K]	Pressure [atm]	Ignition delay [us]
930	**3.586**	**7912**
930	**3.534**	**13090**
...	...	...
938	3.651	7723
938	3.535	6973
938	3.52	6133

this boundary. Conversely, if the ontology lacks properties or classes that are incongruent with experiment uncertainty, the accuracy of embedding will have a negligible impact on the prediction of uncertainty links [44].

In particular, the ontology used for the scenario of this work accounts for:

- *Experiment*. It is a unique identifier of the experiment.
- *Author*. It is a unique identifier associated with the author who publishes the experiment. This entity class can be extended with other information, such as journals or affiliations.
- *Performance*. It is a performance index measuring the similarity between the experimental and simulated data, from 0 to 1, where 1 is the best performance. The simulated data are generated by a model used as a reference. In this study, the possible performance values are discretized equally in ten parts from 0 to 1.
- *Year*. It is the publication year of the experiment.
- *Target*. It is the subject of the experiment. In the general case, an experiment can have multiple properties under observation. It is possible to represent such a case in the KG by adding a new experiment entry for each subject where all the other metadata are unchanged except for the performance and the uncertainty.
- *Type*. It is the typology of the experiment.
- *Instrument*. It is the instrument used for the experiment.

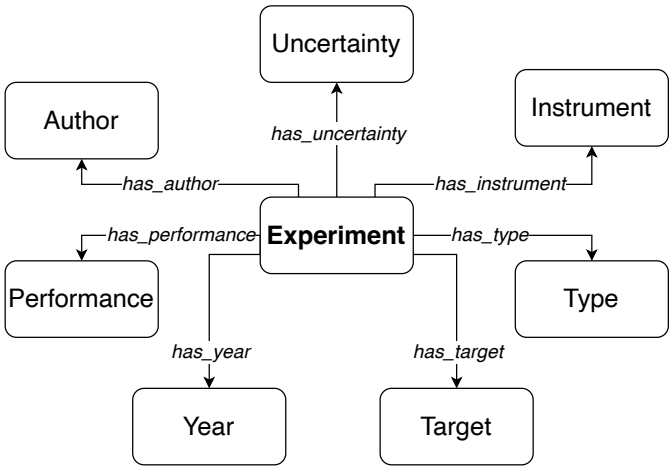


Figure 6.2: *Representation of the metamodel of a typical experiment using a KG.*

- *Uncertainty*. It is the (relative) uncertainty of the experiment if it is provided. The possible uncertainty values are equally discretized in ten parts from 0 to 1. 0 implies that it is not possible to determine the uncertainty.

6.2.2 Knowledge Graph Embedding

KGE generates a representation of a KG in a low-dimensional space of size k, such that entities with similar semantic meanings have close embeddings [200].

The embeddings differ in terms of the following characteristics known as an embedding model: scoring function, representation space, encoding models, and any other additional information that should be included in the computation [201, 202].

The most common class of embedding models uses Euclidean spaces, such as the precursor TransE [200] and the more recent RotatE [203].

TransE interprets the embedding of the entities and the relationships as a translation of the Euclidian space. The scoring function is in Equation (6.1) where $\bar{h}, \bar{r}, \bar{t} \in \mathbb{R}^k$ are the head h, relation r, and tail t embeddings of a triple (h, r, t). The score function, in the case of TransE, measures how distant the vector representing the tail $\bar{t}$ of a triple from the vector of the head $\bar{h}$ plus (vector sum) the vector of the relation $\bar{r}$.

The training procedure minimizes the loss over a training set of triples T as in Equation (6.2) within a number of given epochs.

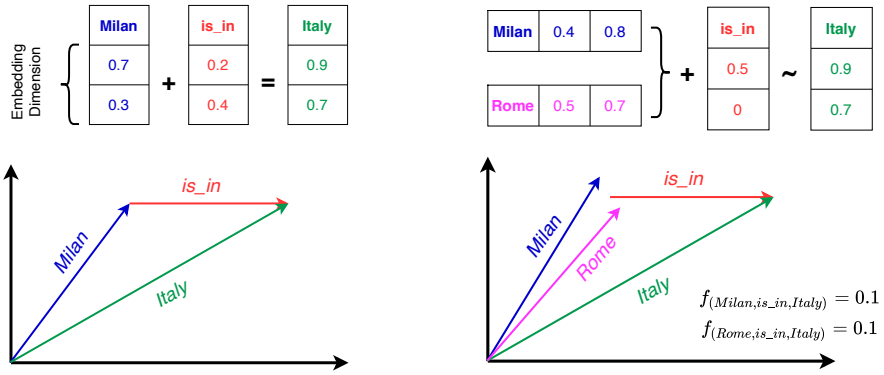


Figure 6.3: *Two step of embedding procedure of a KG. First in Figure 6.3a is shown a possible embedding of a triple in the case of TransE. Then in Figure 6.3b is depicted how the embedding is computed for multiple triples, minimizing the loss.*

Figure 6.3 illustrates a possible embedding representation for two entities and a relation in the case of TransE with an embedding dimension equal to 2.

$$f_{(h,r,t)} = \bar{h} + \bar{r} - \bar{t} \tag{6.1}$$

$$Loss = \sum_{\forall (h,r,t) \in T} f_{(h,r,t)} \tag{6.2}$$

This thesis adopts RotatE. It is a good trade-off between computational complexity and the accuracy of the representation [201]. It is capable of representing, unlikely TransE, complex relational properties such as inversion (e.g., child and parent), symmetry (e.g., marriage), and composition (e.g., my parents' parents are my grandparents). Similarly to TransE, RotatE projects the entities and relations into a complex space. The tail entity is reachable from the head entity by a rotation defined by their relationship. Formally, given a triplet (h, r, t), RotatE learns the embedding $\bar{h}, \bar{r}, \bar{t} \in \mathbb{C}^k$ such that is satisfy the mathematical relation $\bar{t} = \bar{h} \circ \bar{r}$, where $\circ$ denotes the Hadamard product.

Link Prediction

The embedded representation of a KG can be used for Machine Learning (ML) tasks such as Link Prediction (LP), where are inferred the missing

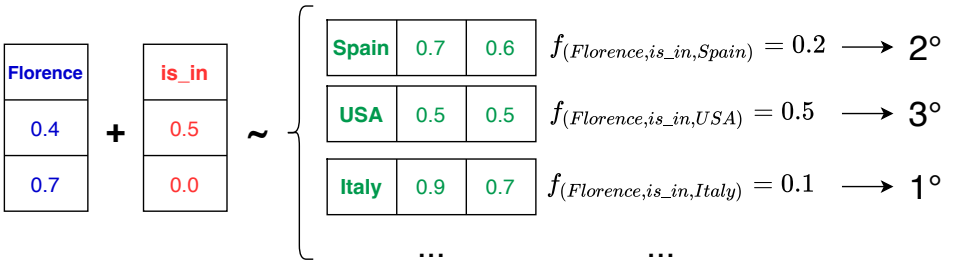


Figure 6.4: *Example of link prediction of the missing entities in a triple. In this case given the head $Florence$ and the relationship is_in, the embedding model predict the missing tail with the entity $Italy$ since it has the lowest score, i.e., it has the highest rank.*

facts. In this work, LP is used to predict the missing uncertainty of an experiment, illustrated by a set of entities representing an uncertainty value, given the embedding of an experiment and the relationship "has_uncertainty".

LP can complete a KG triple when it is missing, one at a time, the head, the relation, or the tail. LP completes a triple with existing entities or relationships during the embedding model's training. For example, when the tail entity is absent in a triple, the process involves the initial gathering of embeddings for the head and the relation. Following this, embeddings for all potential and semantically meaningful tail entities within the knowledge graph are also collected. For each conceivable combination of the head entity, relation, and potential tail entities, the embedding model's score function is applied to compute a score, and the triples are ranked based on this score. The triple with the lowest score, signifying the minimal distance or error, is identified as the most suitable candidate for completing the triple. Figure 6.4 is an example of LP using TransE as an embedding model. In this case, the tail is missing in the triple $(Florence, is\_in, ?)$.

Evaluation Metric

LP stands as a benchmark for assessing the accuracy of embeddings when applied to a test set of triples denoted as Q. In our case, to measure the ability to forecast the missing uncertainties correctly. Hits@N (as defined in Definition 6.2.2) is a metric to measure such performance.

Definition 6.2.2 (Hits@N)**.** *Hits@N (or H@N) refers to the ratio of correctly predicted triples among the top N predictions generated by the embedding model, as defined following Equation* (6.3).

$$Hits@N = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\begin{cases}1 & if\, rank_{(h,r,t)_i} \leq N \\ 0 & Otherwise\end{cases} \in [0,1] \tag{6.3}$$

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

Definition 6.2.3 (Diff)**.** *$Diff$ measures the average error in predicting uncertainty. It is computed as the average difference between the predicted uncertainty value, which is the top prediction of the embedding model, and the actual uncertainty value.*

$$Diff = \frac{\sum_{\forall t \in V} |M_1(t) - R(t)|}{k} \tag{6.4}$$

Given a set of triples V ($|V| = k$), $Diff$ computes, on average, for each triple $f = (h, r, t) \in V$, the average difference between the model's first top prediction $M_1(f)$ and the actual value $R(f)$. This metric assesses whether the embedding model effectively captures the intricate relationships between experiment metadata and uncertainty. For example, consider a triple f where the correct uncertainty value is $R(f) = 0.5$ for a given experiment, and the embedding model's first top prediction is $M_1(f) = 0.4$. While this prediction is incorrect, it is noteworthy that the model's predicted value is relatively close to the actual value. This suggests that the embedding model is learning to predict uncertainty with reasonable accuracy despite misprediction.

6.2.3 Uncertainty Prediction

Uncertainty strictly links the DQ dimensions of completeness, consistency, and accuracy during data profiling. In experimental domains, uncertainty is another experiment metadata; therefore, the scientific repository is "more complete" when the uncertainty information is available. Conversely, when two experiments share identical conditions but differ in authorship, and both lack uncertainty information, several scenarios emerge: either they both report the same observations, both report inaccurate data, or one of them is erroneous. Thus, the presence of uncertainty data holds significant importance in terms of ensuring consistency between experiments and, likewise, in preserving accuracy. Indeed, uncertainty provides a margin of error that facilitates a more accurate evaluation of whether an experiment aligns with simulated data, accounting for some level of deviation. This extends beyond a simple point-to-point comparison of values.

This section outlines the comprehensive methodology for predicting missing experimental uncertainties using KGE. The approach is delineated through the following steps, also represented in Figure 6.5. The methodology starts after collecting a scientific repository of experiments and the ontology for their description. In this study, uncertainty is considered a metadata of the experiment. (i) Experiment profiling is carried out for two primary objectives. Firstly, to assess the data quality level, thus ensuring the highest possible quality for subsequent KG generation. Secondly, to quantify the range and diversity of uncertainty values. (ii) KGE models typically cannot predict continuous values directly. To address this limitation, a finite yet representative and comprehensive set of potential uncertainty values (referred to as *buckets*) is selected based on the specific application domain. The input dataset's uncertainties are then transformed using a bucketization process, grouping similar or close values into the same bucket. (iii) After separating experiments with uncertainty from those without, (iv) a KG of the input experimental dataset is constructed following the provided ontology (as detailed in [144]). Subsequently, training, validation, and test sets are randomly generated, with allocations of 80%, 10%, and 10% of the original dataset, respectively. (v) An embedded representation of the KG is then learned. (vi) The embedded model is subjected to validation, and if the results meet satisfactory criteria, (vi) the embedded model is employed to predict missing uncertainties, with the task primarily centered around link prediction.

The upcoming sections detail the scenarios carried out to examine and affirm the validity of the proposed methodology for predicting uncertainty.

The purposes of the scenarios are the following:

1. Set a baseline against which to compare the performance of the other scenarios.
2. Determine whether it is possible to learn how to predict uncertainty values when it systematically depends on another experiment metadata.
3. Understand how complex the relationship between uncertainty and experiment metadata can be to predict correctly the uncertainty.

6.2.4 Results

Each scenario utilizes the ontology presented in Figure 6.2 as a foundation for constructing the KG, incorporating properties relevant to the specific investigation. These scenarios are intentionally crafted using synthetic data,

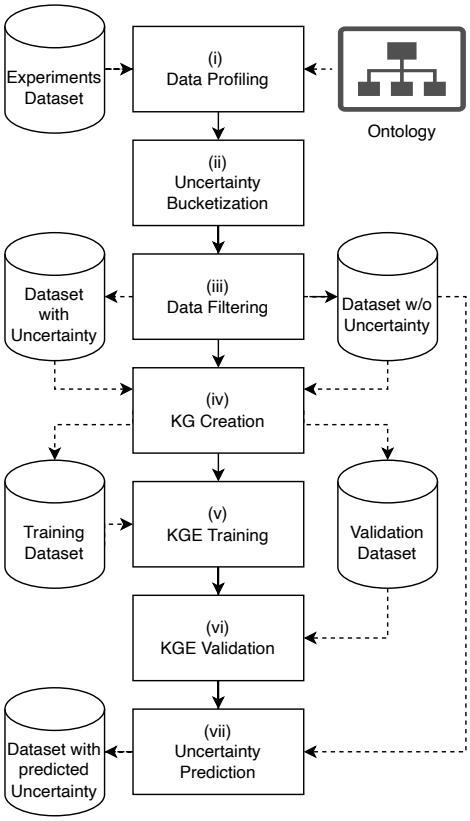


Figure 6.5: *Steps of the proposed methodology to predict the missing uncertainties of experimental data.*

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

Table 6.2: *The number of different values for each entity. "Exp." stands for experiment, "Inst." for instrument, "Uncert." for uncertainty, and "Perf." for performance.*

making them highly suitable for conducting parametric analyses. Furthermore, they offer the flexibility to test the proposed methodology across varying KG sizes, measured by the number of experiments and, consequently, triples, while considering the associated training costs. These scenarios conduct tests with $1,000$ experiments, generating a total of $7,000$ random triples while ensuring compliance with domain constraints among entities. Each experiment was randomly associated (with uniform probability distribution) with entity types present in the ontology, maintaining the appropriate relationships and feasible entity values. For instance, in our application context, not all instrument types may be applicable to every experiment type. To investigate the methodology's robustness and its independence from the number of triples, tests with a larger dataset are performed, involving $50,000$ experiments generating $350,000$ triples. These scenarios maintained a proportional distribution of entities for each typology. Notably, the results indicate that the methodology's performance remains consistent regardless of the number of triples. For reference, Table 6.2 provides information on the number of possible values for each entity type.

The training process for the KGE model is repeated five times for each scenario. Numerical results, referred to as "prediction performance" hereafter, are calculated as the arithmetic average of the outcomes from these five test cases. The list of triples describing the KG for each scenario is randomly divided into three datasets: training, validation, and test, constituting 80%, 10%, and 10% of the total triples, respectively. The settings for the embedding model remain consistent across all test cases. Specifically, RotatE is employed as the embedding model with a fixed embedding dimension of 64. The maximum number of training epochs is set at $15,000$, with an early stopping mechanism based on the $H@3$ score computed every 500 epochs over the validation dataset. Early stopping is triggered when no improvement is observed for three consecutive steps, with a minimum delta of $5\mathrm{E}{-}03$.

The scenarios are assessed in two ways: H@N (Equation (6.3) assesses the prediction capabilities only on the link prediction task for the relationship that connects the experiments to the uncertainty entities. $Diff$ (Equation (6.4)) evaluates the average error in the mispredictions.

Baseline

The KG is generated in the baseline scenario according to the ontology outlined in Figure 6.2. During this generation process, consistency rules are enforced, ensuring that experiments attributed to the same author have plausible publication years within the author's range of activity years. In

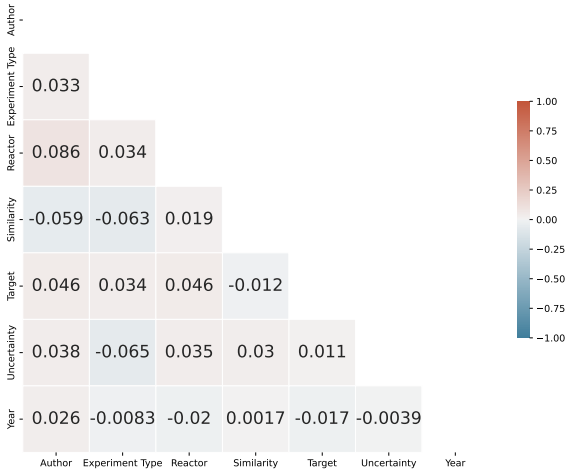


Figure 6.6: *Metadata correlation (Heatmap matrix) in the "Baseline" scenario.*

this scenario, the uncertainty of each experiment is randomly assigned a value between 0 and 1. The correlation heatmap displayed in Figure 6.6 reveals no discernible correlation among the ontology properties. As anticipated, the embedding results align with these expectations. The model is tasked with predicting the correct uncertainty value for experiments from a pool of 11 possible values. The results presented in Table 6.3 demonstrate that the predictive performance does not significantly exceed the theoretical limit since there is no systematic relationship between uncertainty and other metadata. For instance, the metric $H@5$ measures the percentage of times the correct value is found among the top five predictions made by the embedding model. Given the availability of 11 possible values, the probability that the correct value falls within the top 5 predictions is approximately 0.454, which closely aligns with the value of $H@5$ observed in the model. Similar observations apply to the results for $Diff$. These outcomes underscore the intuition that when no discernible pattern exists between uncertainty and experiment metadata, it becomes challenging to learn how to predict missing experimental uncertainties. This suggests that uncertainty may be contingent on other, potentially more complex factors that are not explicitly represented in the knowledge graph.

Scenario 1

Previously, the baseline scenario illustrated that when uncertainty lacks a discernible connection to any experiment metadata or ontology property,

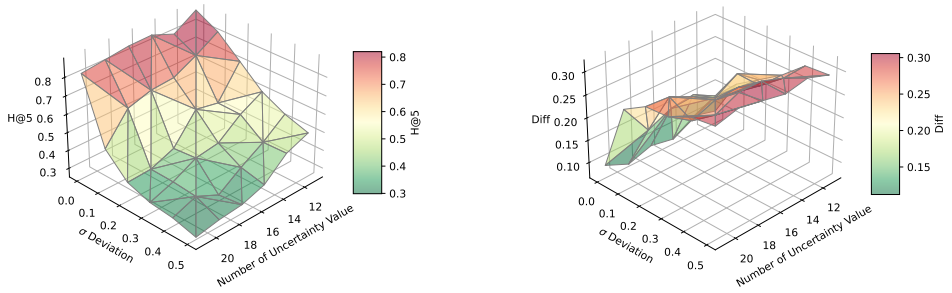
	H@5	H@3	H@2	H@1	Diff
Mean	0.465	0.170	0.142	0.087	0.38
Median	0.479	0.167	0.142	0.073	0.38
Max	0.490	0.177	0.142	0.115	0.39
Min	0.427	0.219	0.167	0.073	0.37
St.Dev.	0.033	0.021	0.005	0.024	0.01
Var.	0.001	0.001	0.001	0.001	0.00

Table 6.3: *Embedding model performance in the "Baseline" scenario.*

the best performance achievable by the embedding model is akin to random guessing when predicting missing uncertainty values. However, the current scenario explores whether learning and predicting uncertainty is possible when a systematic relationship exists between experiment metadata and uncertainty. In this scenario, uncertainty values are determined based on the values of other experiment metadata. More formally, given $X1$ an experiment metadata and $X1 = \{X1_1, ..., X1_n\}$ the possible n value that $X1$ can assume in a domain. Given $U = \{U_1, ..., U_k\}$ where $U_j \in [0, 1]$ the k uncertainty values present in the knowledge graph. Specifying the relationship $\forall X1_i \in X1 \rightarrow U_j \in U$ is needed. In this scenario, the relationships are initially selected at random from a uniform distribution and remain constant throughout the knowledge graph generation process. As a result, each potential value of $X1_i$ consistently corresponds to the same uncertainty U_i. However, this strict association is unlikely to hold true in a real-world scenario. A parametric analysis is conducted to account for this variability. This analysis involves increasing the cardinality of U while introducing a *run-time* modification (i.e., during knowledge graph generation) in random deviations with bounded positive or negative values akin to random noise. Specifically, the relationship between $X1$ and U is adjusted from $X1_i \rightarrow U_j$ to $X1_i \rightarrow U_j \pm \sigma$, where σ takes on values from the set $0.0, 0.1, 0.2, 0.3, 0.4, 0.5$. This analysis accounts for the more realistic scenario where the association between $X1$ and U may exhibit some degree of variability.

The outcomes of the parametric analysis, as depicted in Figure 6.7, are assessed based on the $H@5$ score (Figure 6.7a) and the $Diff$ measure (Figure 6.7b). The results reveal that the KGE model consistently delivers strong performance, regardless of the number of uncertainty values, when the absolute deviation is tightly bounded. This indicates the model's ability to discern and learn systematic patterns between experiment metadata and uncertainty under such controlled conditions.

Conversely, when the absolute deviation is set to 0.5, replicating the



(a) **(b)**

Figure 6.7: *$H@5$ (Figure 6.7a) and $Diff$ (Figure 6.7b) 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.*

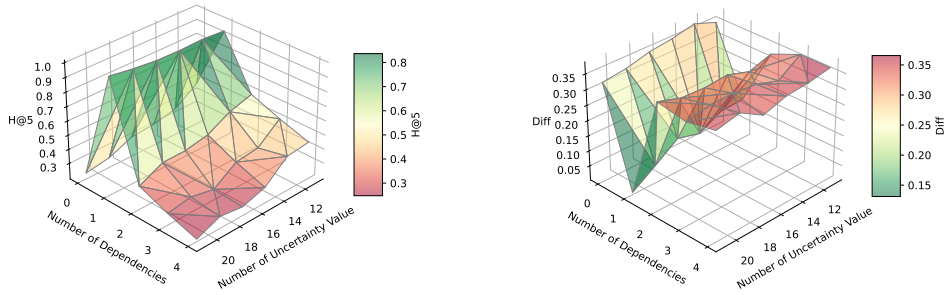
conditions of the baseline scenario, the results align with the expectations set by the baseline scenario. Given that uncertainty values are constrained within the range of $U_j \in [0, 1]$, introducing an absolute deviation of 0.5 from the mean uncertainty value of 0.5 effectively allows all possible values to be associated with the same metadata. Consequently, no distinguishable relationship exists between metadata and uncertainty in this scenario.

Scenario 2

The previous scenario demonstrated that KGE can effectively predict experiment uncertainty when a systematic relationship exists between experiment metadata and uncertainty, even in the presence of randomness. This scenario wants to explore whether this methodology remains effective when uncertainty depends on an increasing number of metadata factors, thereby investigating the complexity of the relationship between experiment metadata and uncertainty values.

To evaluate this, a parametric analysis is conducted in which both the cardinality of U and the number of metadata dependencies are expanded. In the previous scenario, the relationship between experiment metadata and uncertainty values was expressed as $X1_i \rightarrow U_j$. However, in this case, the general relationship is extended to $(X1_i, ..., X4_m) \rightarrow U_j$, indicating that uncertainty values can depend on $0, 1, 2, 3$, or 4 experiment metadata values.

Figure 6.8 presents the result of the parametric analysis. The results are evaluated based on the $H@5$ score (Figure 6.8a) and the $Diff$ measure (Figure 6.8b).



(a) **(b)**

Figure 6.8: *$H@5$ (Figure 6.8a) and $Diff$ (Figure 6.8b) 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.*

As observed previously, when there is no dependency between experiment metadata and uncertainty values, the performance closely resembles that of the baseline scenario. When dependency exists for only one metadata value, it aligns with the patterns observed in the previous scenario. However, as the number of metadata dependencies increases, the embedding performance degrades, albeit still remaining significantly better than the baseline results. This trend holds true regardless of the number of possible uncertainty values. These findings suggest that the KGE model enhances performance over the baseline scenario, although not to the same degree as in simpler cases.

Scenario 3

In this real-world scenario, the same methodology is applied to a dataset of chemical kinetics data accessible within the SciExpeM data ecosystem [1]. A subset of 440 experimental data points, each with uncertainty information, has been collected. The corresponding knowledge graph encompasses a total of 11,000 triples, involving six different values of uncertainty. Additional information regarding the number of possible values for each entity is provided in Section 6.2.4.

Similar to the earlier scenarios, the results presented in Table 6.5 indicate that the embedding model can predict missing uncertainties, even though the task is somewhat more manageable. This is because the number of distinct uncertainty values is reduced to six, as opposed to the 11 values considered in the general ontology setting for these scenarios. Con-

[1] `https://sciexpem.polimi.it`

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

Table 6.4: *The number of different values for each entity. "Exp." stands for experiment, "Inst." for instrument, "Uncert." for uncertainty, and "Perf." for performance.*

	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
St.Dev.	0.02	0.02	0.02	0.01
Var.	0.001	0.001	0.001	0.00

Table 6.5: *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.*

sequently, measuring performance indexes beyond $H@3$ becomes impractical, as the model needs to make predictions from a more limited set of potential values. Nevertheless, the obtained results remain promising and encouraging.

In a broader context, the methodology described can be applied to generate knowledge graphs in various applications where an ontology can be defined. The complexity of the relationships between entities within the knowledge graph directly influences the learning potential of KGE techniques. The approach outlined here can be employed to investigate whether a dependency exists between ontology properties and uncertainty. In such cases, the embedding model can evaluate the feasibility of predicting uncertainty and provide insight into the possible values of such predictions.

Experimental uncertainty is fundamental for the DQ profiling of scientific data, as well as for other predictive model development tasks in which the experiments drive the model development. 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. Since having multiple observations is rare in practice, this contribution proposes a new methodology to predict the missing uncertainty of experimental data. It leverages only the available information and extracts hidden patterns between the experiment metadata and the uncertainty values, categorized in n different classes, using the machine-learning link prediction task. To do so, an embedded representation of the KG that

corresponds to the experiment repository is learned . This methodology is mainly studied with two parametric analyses focused on understanding whether the KGE 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.

6.3 Data Transparency

Data provenance, also known as data lineage, is a field of data management. It is related to the reproducibility and transformation history of data. Data transparency specifies how and what should be represented concerning the various stages that alter data. Provenance metadata is a solution to enhance transparency and trustworthiness, enriching the final dataset by incorporating a set of metadata. This metadata encodes information about the individuals, actions, and entities involved in the data transformation stages. Provenance is suitable for tracking data transformations but also plays a vital role in recording data processing replicability and identifying errors in data-driven applications [204, 205].

The provenance data model outlines the structure of the provenance metadata, defining how to represent entities, activities, and agents and their interactions involved in data creation or transformation. In this work, the W3C PROV data model[2] is employed to design the provenance data model. Adhering to the guidelines set forth by data sheets directives [206], the provenance data model aims to represent only the information essential to the subject under study. Generally, a provenance data model can vary in specificity, altering the verbosity level [205].

The W3C PROV data model relies on three fundamental concepts for representing provenance metadata, namely *Entity*, *Activity*, and *Agent* [207]. An *Entity* is the subject of the provenance; it is something whose evolution is tracked. An *Activity* denotes an action conducted on an *Entity*, resulting in the creation of a new version of itself or another *Entity*. An *Agent* embodies the *Entity* responsible for carrying out a specific action or being associated with a particular entity. The W3C data model, as depicted in Figure 6.9, encompasses a blend of the "data" and "workflow" levels of detail, presenting the requisite provenance metadata to describe the stages within the business process outlined in Figure 5.5. However, it excludes the

[2] https://www.w3.org/TR/prov-dm/

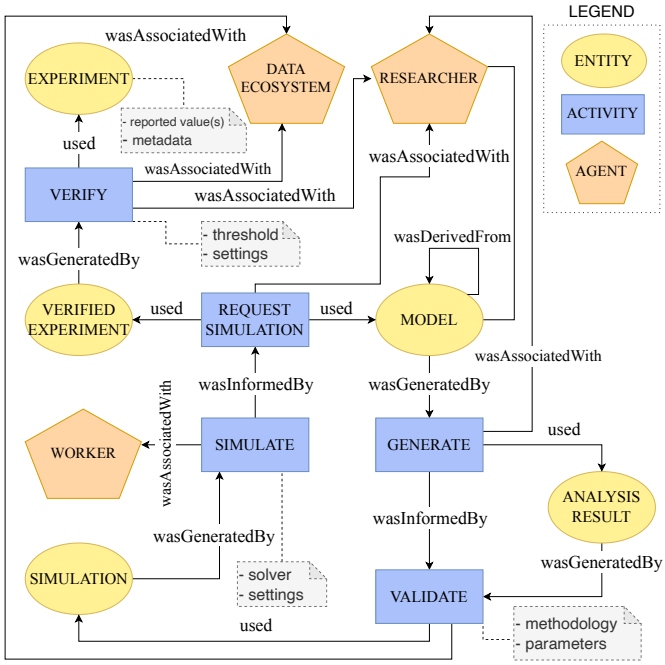


Figure 6.9: *Representing provenance of the model development process in the context of scientific data within a DE with the W3C PROV-Data model.*

procedure related to the external collection of experiments and models. On the other hand, it incorporates novel components introduced in this work's proposed solution. These additions include a verification step designed to enhance data repository quality and the implementation of a coordinator-worker configuration for distributed computational workload management.

Multiple experiments and models are stored in the DE. An *experiment*, before entering the loop of the model development process, is submitted to a validation process. The *Data Ecosystem* and *Researcher* automatically and manually, respectively, *verify* whether a new *Experiment* is compliant with the established quality thresholds in the policies defined by the DE community.

As a result, a *Verified Experiment* is derived and stored within the *Data Ecosystem*. At the first iteration of the loop, a *Model* is given, and the *Researcher requests* the *Simulation* of a set of *Verified Experiments* with the *Model*. At this point, a computational task is assigned to the *Workers* that *simulate* the *Verified Experiment* generating a *Simulation*. Once all the *Simulations* are terminated, the *Data Ecosystem* can perform an analysis of the results, such as model *validation*, thus providing *Analysis Result*,

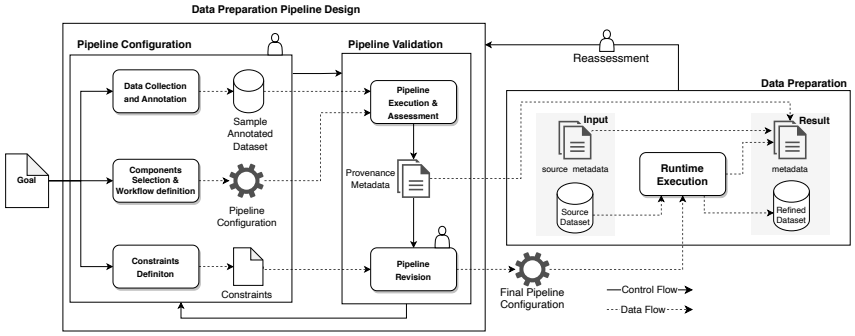


Figure 6.10: *Data preparation pipeline design process*

fundamental for the *Researcher* to *generate* a new, improved version of the *Model* and start the model development process again.

6.3.1 Data Pipelines

High-quality datasets are fundamental to minimizing errors and risks in applications where analytics are the primary resource for supporting decision-making. To achieve this, a comprehensive data preparation pipeline must precede data analysis tasks, as discussed in the next Chapter 7. Data analysts must carefully select various operations to be included in the pipeline, considering multiple aspects. This selection process often requires multiple trials and errors. A few times, it leads to the most effective solution. In this context, data transparency has a double purpose. Firstly, it serves as a repository for different configurations of the data preparation pipeline, incorporating quality and performance indicators. The data analyst then uses such data to modify the design of the pipeline accordingly. Second, to enable data sharing and reusability, the generation of metadata related to the processing phases has been recognized as essential. Automatically annotating the dataset resulting from the pipeline with provenance metadata addresses this issue. For this work, the case study is in the citizen science field. More precisely, social media data. Extracting meaningful insights from social media presents distinct challenges, including isolating pertinent posts, contextual dependence, multimedia elements, and the potential for inadvertent exclusion of informative content via automated filters. Therefore, this section presents the proposed approach with a particular emphasis on the provenance model for these purposes.

The data analyst, given an input dataset, a set of available tools to process it, and a *Goal*, wants to identify the most suitable sequence of steps to

be performed, the configuration of the selected tools, and an assessment of the expected quality of the outcome of the preparation. The *Data Preparation Pipeline Design* is presented in Figure 6.10 and comprises two macro-phases, *Pipeline Configuration*, and *Pipeline Validation*, briefly described in the following.

The data analyst performs the following actions during *Pipeline Configuration phase*. A representative sample is extracted from the initial dataset and annotated the relevance of the data items with respect to the predefined goal (*Data Collection and Annotation*). The data analyst selects, from the *Component Library*, the components that have to be included in the pipeline and their execution order and settings (*Components Selection and Workflow definition*). Data analysts can also express some non-functional constraints related to performance and quality measures of the entire pipeline (*Constraints Definition*). As a result, this phase produces three artifacts: a *Sample Annotated Dataset*, a *Pipeline Configuration*, and a list of non-functional *Constraints*. These objects are provided to the *Pipeline Validation phase*, which aims to evaluate the efficiency and effectiveness of the initial pipeline configuration and improve it, if necessary. In particular, during the *Pipeline Execution and Assessment*, the pipeline is executed on the sample dataset, and the quality of the result is assessed with respect to the sample annotations. Execution information, including involved components and related configurations, and components and pipeline performances, are stored as *Provenance Metadata*. Later, during the *Pipeline Revision*, the results of the pipeline execution in terms of performance and quality are presented to the data analyst through a feedback dashboard. In case of unsatisfactory outcomes, the data analyst might decide to modify the pipeline going back to the Pipeline Configuration phase. The system can also support the data analyst by suggesting enhancements such as component substitution and/or reconfiguration.

When the data analyst is satisfied with the pipeline configuration, the *Data Preparation* phase can be executed on the *Source Dataset*, i.e., the entire data source from which the sample set has been extracted. The data preparation pipeline takes as input such data together with their metadata in order to obtain the *Refined Dataset* that will be used in the data analysis. At the end of this phase, or even at a later moment, it might happen that the data analyst is not satisfied with the obtained results. This could happen for different reasons, e.g, the Source Dataset was not accurately represented by the sample data, and/or the characteristics of data changed over time. In this case, the redesign of the pipeline is needed, and the whole process is reiterated.

The refined dataset generated through the pipeline can be used in a specific analysis but could also be published in a DE. In order to enhance the reusability of the dataset and the transparency of the data preparation phase, the refined dataset is enriched with metadata. Three types of metadata are associated with it: (i) **source metadata**, directly obtained by the data source and associated with the source dataset, (ii) **execution metadata**, collecting information about the pipeline execution on the items of the source dataset; (iii) **provenance metadata**, generated in the configuration phase and capturing the pipeline characteristics.

The availability of such metadata also makes our approach beneficial for the creation of data spaces in which the trustworthiness among partners is a key factor based on ensuring transparency in data preparation and high-quality of the shared data. For example, in data lakes, there is usually a raw data area containing ingested data and a revised data area where clean data sets are stored [208]. The proposed approach can be used to generate clean datasets from the raw data. Moreover, the availability of provenance metadata associated with the refined dataset allows data analysts to understand better if the data preparation tasks performed are suitable for their analysis.

In this work, as in all user-driven data preparation approaches, provenance is central when a series of pipeline steps for preparation manipulate data. For example, in a pipeline that manages scientific data to develop a data-driven model, it is hard to disambiguate which phase or procedure of a long and complex process is responsible for the improvement or deterioration of the model. Provenance, keeping track of each action on the data and on the model, can help in this task and make it possible to replicate scientific results [204]. In general, the data analyst can be motivated to use provenance for multiple purposes such as accountability, reproducibility, or process debugging [205], but in our methodology, it also increases transparency and trustworthiness.

The PROV data model is adopted to specifically track the provenance in data preparation pipelines for social media data as shown in Figure 6.11. Since it must fully track the evolution of the pipeline configuration and it also must enable the reproducibility of the refined dataset generation.

The PROV data model records metadata about the Sample Dataset extraction, the Pipeline configuration, and the execution of the pipeline with a given configuration, as well as the sequence of revisions of the configuration performed by the data analyst.

Following the case study, but without losing generality, a *Sample Dataset* is generated by a *Crawling* action using a specific *Crawler* on a given *Source*. For social media analysis, it is important to keep track of the search

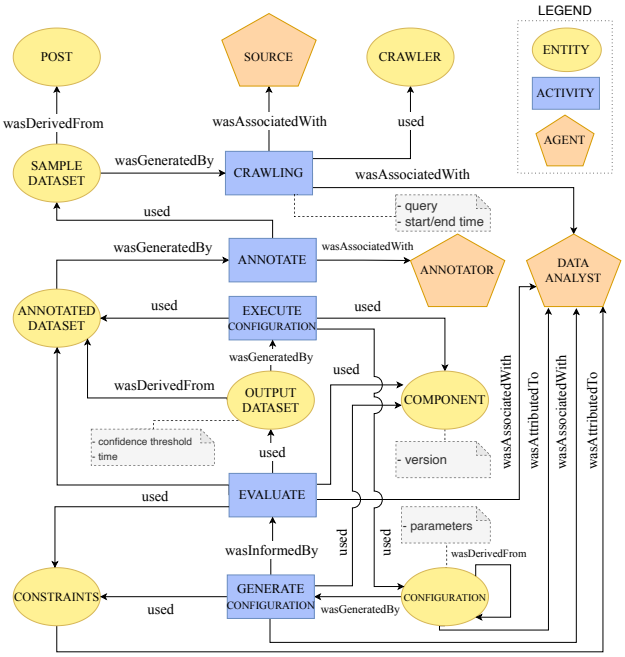


Figure 6.11: *The conceptual provenance data model for social media data preparation pipelines*

parameters such as the query and the time interval, which are associated with the Crawling action, and when the *Crawler* was interrogated to collect data. Even the type of used *Crawler* could have an impact on the final results. For example, Twitter, which could be a *Source* of social media data, has different types of Application Programming Interface (API) (developer, academy, business, etc.), each with a different level of granularity in collecting tweets and thus different final results. For this reason, inside the *Crawler* agent, it is important to specify the crawler version. The resulting *Sample Dataset* is a set of *Post*s. Once the *Sample Dataset* is available, the *Annotator* is responsible for the execution of the *Annotate* action on the collected data, generating the corresponding *Annotated Dataset*. The *Data Analyst* generates also the first initial *Configuration* of the pipeline with the *Generate Configuration* action. A *Configuration* includes *Component*s, their parameters, and the Components' workflow. The *Constraints* defined by the *Data Analyst* for the overall pipeline are recorded. *Execute Configuration* records the *Component* actions using a given *Configuration* to produce the resulting *Output Dataset*; for each *Component* its characteristics are recorded (e.g., component version). The time and cost

of the pipeline execution with a given configuration are recorded in *Execute Configuration*. The *Evaluate* action by the *Data Analyst* is based on the assessment of the *Output Dataset* against the *Annotated Dataset* and *Constraints*. Leveraging the quality performance information of the overall pipeline and of each *Component*, the *Data Analyst* may decide to generate (*Generate Configuration*) a new *Configuration* as an improvement of the previous one (*DerivedFrom* on *Configuration*) or modify the *Constraints* to reach the desired outcome. The actions *Execute*, *Evaluate*, and *Generate* may be iterated until the *Constraints* specified for the pipeline are fulfilled and the *Data Analyst* confirms the final *Configuration*. The whole process is recorded as metadata associated with the prepared dataset, as it describes how the *Configuration* for data preparation was achieved. Hence, they store quality, cost, and time information about the pipeline execution.

6.4 Diversity

Model Validation (Section 7.1) involves assessing how well a model's predictions align with actual experimental data. During this process, the reliability of the predictive model and subsequent analyses is not solely determined by the quantity and quality of experimental data. It is equally important to consider the diversity of the data used, specifically in terms of how comprehensively it covers the domain that the model intends to represent. This discipline is named database coverage (or diversity) [165].

For instance, simply using many data for validation, all describing the same portion of the domain, may not suffice to determine the reliability of the model validation outcomes. Indeed, if a model has not been validated under various environmental conditions, its performance may unexpectedly deteriorate when employed to predict an unexplored (untested) yet physically relevant part of the domain.

A lack of adequate coverage in the dataset, thus limited testing of the generality capabilities of a model, can lead to biased reliability in model validation (Section 7.4) [209]. Ideally, the collection of experiments should be as extensive and varied as possible, against which the model can be validated. On the other hand, knowing the diversity of a database can be leveraged to identify areas of the domain that are inadequately represented by data, thus facilitating the Design of Experiment (DOE) process. This thesis introduces a methodology that utilizes categorical attributes and a multidimensional matrix to represent the diversity of a domain and establish a database coverage index.

The assessment of dataset coverage $\mathcal{C}$ for a domain $\mathcal{M}$ with n attributes,

denoted as $A = A_1, ..., A_n$, is conducted in three step.

First, it is necessary to identify a subset of the domain fields (or attributes) $\{A_1, ..., A_s\} = \hat{A} \subseteq A$, and transform them into *categorical attributes*. A categorical attribute of a domain is a field that can only take a value from a restricted number of options. In this way, any attribute $A_i \in \hat{A}$ can only have d_{A_i} different ordered categorical values (or possible options). If the attribute $A_i \in \hat{A}$ is a continuous numeric field, the minimum (*min*) and the maximum (*max*) value that can be taken by A_i in the domain, and fix t equidistant ticks from the range $[min, max]$ and associate the value of the attribute to the closest tick. Instead, suppose the possible values of an attribute are not continuous but with high cardinality. In that case, identifying a subset of the possible values leveraging a hierarchy among them or using the bucketization: similar values are associated with the same bucket [209]. Given an entry r of the domain $\mathcal{M}$ regarding an attribute $A_i \in \hat{A}$, it has a corresponding value of $v_{A_i,r} = (v_{1,i}, ..., v_{d_{A_i},i})$ for the attribute A_i where $v_{i,j} = 1$ if r has the corresponding categorical value for the attribute A_i otherwise is 0. In this way, it is possible to register an array field of the model where an entry can assume multiple categorical values for the same attribute. The notation $v_{A_i,r}[k]$ to denote the k-th value of the attribute A_i with $k \in [1, d_{A_i}]$ for the entry r.

In the first step are selected the most significant subset of the domain's attributes, denoted as $A_1, ..., A_s = \hat{A} \subseteq A$. Then, they are transformed into "categorical attributes." A categorical attribute within the domain is a field that can only assume values from a limited set of options. Therefore, for any attribute $A_i \in \hat{A}$, it can have only d_{A_i} distinct ordered categorical values or possible options. If attribute $A_i \in \hat{A}$ is a continuous numerical field, it is computed the minimum (*min*) and maximum (*max*) values it can take within the domain. Then, this range is divided into t equidistant intervals and associates the attribute's value with the closest interval. Conversely, if the possible values of an attribute are not continuous but possess a high cardinality, a subset of these values is identified, using a hierarchy strategy or employing bucketization, grouping similar values into the same bucket [209]. For a given entry r within the domain $\mathcal{M}$ associated with an attribute $A_i \in \hat{A}$, it is assigned a corresponding value of $v_{A_i,r} = (v_{1,i}, ..., v_{d_{A_i},i})$ for that attribute. Here, $v_{i,j} = 1$ if entry r possesses the corresponding categorical value for attribute A_i; otherwise, it is set to 0. Consequently, this approach allows the recording of a database entry that can assume multiple categorical values for the same attribute. Thee notation $v_{A_i,r}[k]$ represents the k-th value of attribute A_i, with k ranging from 1 to d_{A_i}, for entry r.

The second step involves constructing a multidimensional space that

represents the coverage of our database across the set of attributes $\hat{A}$, where the cardinality of this set is denoted as $|A_s| = s$. Each attribute $A_i \in \hat{A}$ defines a dimension in this space, and the dimension's size is determined by d_{A_i}. Subsequently, the multidimensional space is translated into a multidimensional matrix known as the coverage matrix $\mathcal{C}_\mathcal{M}$ with dimensions $d\mathcal{C}_\mathcal{M} = dA_1 \times ... \times d_{A_s}$. This matrix-based approach allows to investigate the database diversity with different levels of granularity.

Finally, after initializing all the matrix cells to 0, it is necessary to iterate through every entry r in the model $\mathcal{M}$ and for every possible combination of categorical attribute values. The coverage matrix is updated following Equation (6.5) only if the specified condition expressed by Equation (6.6) holds true for entry r when $i_m \neq 0$, where m ranges from 1 to s. In other words, by employing a technique known as "bucketization," the matrix is filled with the number of experiments available within a given domain region corresponding to one or more cells (or boxes) of the multidimensional matrix. Bucketization involves grouping similar values into the same bucket. For instance, if a dimension has been divided into buckets with values of $0, 5, 10, 15$, data points with values for that dimension of 2 are associated with bucket 0, while those with values of 3 are associated with bucket 5, and so on. The definition of the buckets can be tailored to the specific characteristics of the domain.

Ultimately, the database coverage can be assessed using a threshold t, which is defined as the ratio between the number of different cells that have at least t associated experiments or data (denoted as $|cells(t)|$) and the total number of cells in the matrix M, as expressed in Equation (6.8).

$$\mathcal{C}_\mathcal{M}[i_1, ..., i_s] \mathrel{+}= 1 \tag{6.5}$$

$$v_{A_1,r}[i_1] == ... == v_{A_s,r}[i_s] == 1 \tag{6.6}$$

The outcome is a density matrix that represents the coverage of our database with respect to a specific set of categorical attributes. A database coverage index can be defined as follows. By examining all the entries r contained within the dataset $\mathcal{D}$, counting the non-empty cells—those cells with values distinct from zero—and normalize this quantity based on the total number of cells, as illustrated in Equation (6.7).

$$\mathcal{C} = \frac{\sum_{i \in [1,d_{A_1}],...,k \in [1,d_{A_s}]} 1, \quad if C_\mathcal{M}[i, ..., k] \neq 0}{d_{\mathcal{C}_\mathcal{M}}} \in [0, 1] \tag{6.7}$$

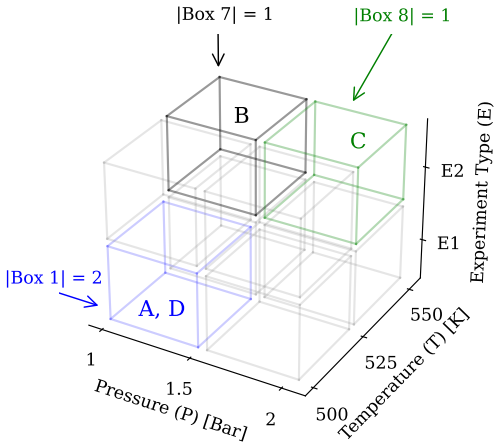


Figure 6.12: *Visualization of the database coverage.*

Experiment	T [K]	P [bar]	E	Box
A	508	1.2	E1	1
B	540	1.2	E2	7
C	537	1.1	E2	8
D	520	1.3	E1	1

Table 6.6: *The experimental data set used for the running example in Figure 6.12.*

$$C(t) = \frac{|cells(t)|}{|M|} \tag{6.8}$$

One or multiple multidimensional matrices can be employed to depict the diversity of a database in various scenarios.

Figure 6.12 illustrates a “small world” example where the dataset’s diversity is measured. The dimensions that characterize the diversity are temperature (T), pressure (P), and experiment type (E). In this case, each experiment in the dataset is characterized by a value property (dimension) that positions it within a specific domain region. The temperature dimension ranges from 500K to 550K, the pressure dimension from 1 bar to 2 bar, and the possible experiment types are ’E1’ and ’E2’. In the case of numerical dimensions like temperature and pressure, each dimension is equally divided into two buckets. Categorical properties, such as ’experiment type’ in this example, determine the number of buckets themselves. As depicted in Figure 6.12, this configuration results in eight cells in the

Box	[T, P, E]	Cardinality
1	[(500, 525), (1.0, 1.5), E1]	2
2	[(500, 525), (1.5, 2.0), E1]	0
3	[(500, 525), (1.0, 1.5), E2]	0
4	[(500, 525), (1.5, 2.0), E2]	0
5	[(525, 550), (1.0, 1.5), E1]	0
6	[(525, 550), (1.5, 2.0), E1]	0
7	[(525, 550), (1.0, 1.5), E2]	1
8	[(525, 550), (1.5, 2.0), E2]	1

Table 6.7: *The results of bucketization for the running example in Figure 6.12 using as data set Table 6.6.*

matrix (or boxes) that partition the domain. Table 6.6 presents the experimental dataset used, which includes four experiments, each characterized by its temperature, pressure, and experiment-type features. Through the process of bucketization, each experiment is associated with a specific box, as shown in Table 6.7. Finally, the number of boxes with at least one associated experiment is three. Consequently, the coverage index in this case is $C(1) = 3/8 \approx 38\%$, and if the threshold is set to two, the coverage index becomes $C(2) = 1/8 \approx 13\%$.

Model validation is a central phase in the model development process. To properly perform this activity, this chapter has presented why and how it is necessary to use quality data, the correct pipeline to prepare them, a diverse dataset, and how to guarantee the reusability of the data to enhance trust and engagement through data transparency. Moreover, to properly validate a scientific predictive model, it is necessary to know the experimental uncertainty and predict it when missing. Once the model validation results are properly generated, collected, and distributed, they can be used for the following analyses and improvements described in the following chapter.

CHAPTER 7

Model Evaluation and Improvement

The previous Chapter 6 debates the importance of data preparation in every data-driven application and, in particular, for scientific data. Chapter 3, and later in Chapter 5, explain how the adoption of a Data Ecosystem (DE) changes the business process for the development of a scientific predictive model. Such change not only speeds up the overall process, which was mainly a manual procedure, but includes new steps that improve the reliability and final quality of the predictive models, with, for instance, data preparation, data transparency, and data science techniques. Moreover, Chapter 3 and Section 5.3 describe how the validation and the following analysis phases are central to the predictive model development process, in particular, to improve the model's predictive capabilities. However, as stated in Chapter 2, there is no standardized procedure for doing so, and a ubiquitous, fair, and transparent procedure is necessary to make, for instance, different predictive model performances comparable and automate the model improvement stages [110]. Therefore, Section 7.1 suggests a systematic and automatic model evaluation procedure. Such a procedure first objectively assesses the model's predictive performance and then automatically and systematically studies the model's predictive capabilities, suggesting future improvements.

Section 4.1 illustrates that simulations are particularly computationally expensive. In many practical situations, it raises the necessity to develop a surrogate model (or metamodel), usually with a black-box strategy [34]. Therefore, it is necessary to take a step forward in the context of predictive model generation. Section 7.2 proposes a new adaptive sampling algorithm that selects the smallest and most representative set of training data that can be employed to create a metamodel for faster predictions and, thus, reducing the expensive resources needed for training. In addition, adaptive sampling algorithms can also be used for Design of Experiment (DOE), identifying the domain settings that lack representative data.

Similarly, Section 7.3, centered around the Chemical Reaction Neural Network (CRNN), studies the application of this particular Neural Ordinary Differential Equation (NODE) that combines the generalization capabilities of black-box neural networks with the integration of chemical-physical law, typical of white-box predictive models, in the learning procedure. This preparatory study aims to develop a predictive model for hydrogen data integrating element conservation in the Neural Network (NN).

Finally, Section 7.4 discusses the ethical implication, hazard, and mitigation regarding the predictive model development process.

7.1 Model Evaluation

Even though scientific predictive models have been developed for decades, no unique and detailed evaluation methodology exists. The model validation (or assessment) procedure links the development of a predictive model to the experimental data. It quantifies the predictive model's performance by comparing its predictions with corresponding experimental measurements. Traditionally, this comparison has been conducted manually through a graphical approach, where experimental and simulated data are plotted together in the same figure, and researchers assess whether the model's predictions are sufficiently accurate to consider the model acceptable [175]. Even if model validation is a "poorly posed problem" [175], this approach has two limitations: (i) It lacks objectivity because different individuals can perceive the evaluation differently. (ii) Manual validation is not extensive due to its time-consuming nature, which is highly dependent on the availability of human resources. Consequently, it is only possible to modify the model occasionally since each change requires re-validation of the model to check the impact of the modification on the predictive performance. This problem is known as the "short blanket" dilemma [120]. Additionally, a manual validation fails to extract systematic insights about the

model's behavior from large datasets, which could be central in discovering recommendations for enhancing the model [120]. Introducing a numerical, hence objective, validation procedure addresses the first challenge by delivering an impartial assessment of the model's performance. This section, therefore, proposes the following model *evaluation* procedure that follows the principles of *objectivity* and *systematicity*. It involves three main phases (see Figure 7.1): data assessment, model validation, and model analysis.

As discussed in the previous Chapter 6, the data assessment phase is a fundamental step. In this phase, various data characteristics are evaluated since they directly influence the quality of the outcomes. It quantifies the quality and diversity of the validation set of experimental data chosen for the following model validation and analysis. It is crucial that the collection of experimental data is as extensive, diverse, and high-quality as possible to mitigate the risk of overfitting the model to the selected data or providing erroneous information in the subsequent phases. This phase can be partially conducted during the model evaluation procedure and partly supported by automatic data quality control procedures defined within the management of the DE. During the model validation phase (Section 7.1.1), the similarity, represented as a score, between the experimental data and the model predictions (or simulated data) is quantified. The similarity score, or index, is a measure of how closely the simulated data aligns with the experimental data, and it typically ranges from 0 to 1, where 1 indicates perfect similarity. Finally, the model analysis, as described in Section 7.1.2, utilizes the validation results and data science techniques to extract insights about *which* aspects of the model perform inaccurately, *where* these inaccuracies occur, *why* they occur, and *how much* they affect the model's performance.

A DE is required to implement such an automatic procedure to elaborate *big data* and promote collaboration within the scientific community. It not only helps streamline and significantly reduce the time and human-related errors involved but is also a prerequisite for successfully implementing this automated approach. Table 7.1 provides an overview of the available techniques, categorized into quantitative and qualitative approaches. These techniques regard both model validation tools and analysis technologies, which serve the purpose of systematically understanding the model's behavior. Validation and analysis methods are further classified into coarse and fine-grain categories. Coarse-grain methodologies are capable of summarizing multiple aspects and handling large volumes of data, while fine-grain approaches involve deeper and more pinpointed investigations.

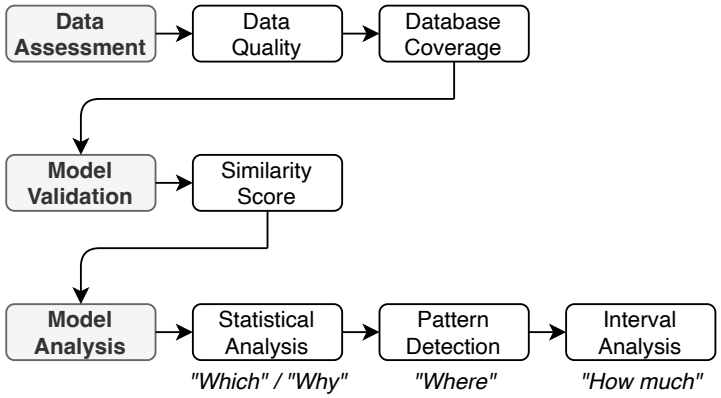


Figure 7.1: *The proposed model evaluation methodology.*

		Validation		**Analysis**
		Qualitative	*Quantitative*	*Quantitative*
Grain	*Coarse*		Trend Score	Pattern Detection Statistical
	Fine	Visualization	Point-Wise Score*	Interval

Table 7.1: *Techniques used to validate and analyze a model. * denote a tool that is present in the literature, but it is not used in our approach. Each tool is quantitative or qualitative and provides detailed (Fine grain) or general information (Coarse grain).*

7.1.1 Validation

The validation phase regards the quantification of the predictive model performances. To be *objective*, the procedure employs a quantitative approach that, by measuring the similarity between the trend (*Trend* score) of the experimental data points as a whole against the corresponding simulated data, provides a synthetic index of the model's performance. The similarity score is computed for each experimental and simulated data pair. Once this operation is concluded, the average of all similarity indexes provides a synthetic overview of the predictive model performance.

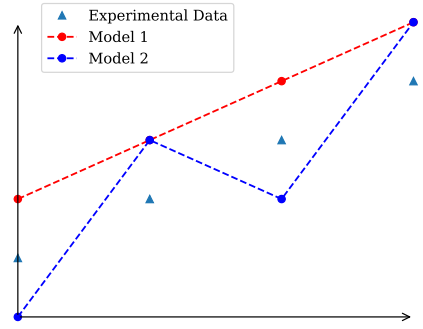
As introduced in Chapter 2, qualitative and quantitative are the two macro families of model validation techniques used in the literature to compare experiments to simulations (Table 7.1). Some of these techniques rely on *Cubic spline interpolation* to derive a continuous function from a discrete data set like in parametric experimental measurements. Cubic spline interpolation defines piecewise function using third-order polynomials, which pass through the given set of data points [210]. Visualization is a subjective, and thus qualitative, comparison of the experiments against the simulated data. Based on their expertise, the researchers evaluate the predictive model performance without quantifying the prediction quality. Moreover, different experts could have dissimilar opinions on the same experimental and simulated pair comparison.

Point-wise approaches define a set of score functions to measure quantitatively the similarity between the experimental and simulated dataset evaluating the error point by point. These approaches are fast to compute, but they do not consider that the points are stand-alone but belong to a chemical-physical measurement trend of phenomena that could lead to misleading results. In Figure 7.2a an example of this pitfall: even if the trend of the simulated data point of *Model 1* is quite different from *Model 2*, the point-wise error of the models when computed against the experimental data is the same. In such a family of scores, one of the most frequently used is the following definition 7.1.1.

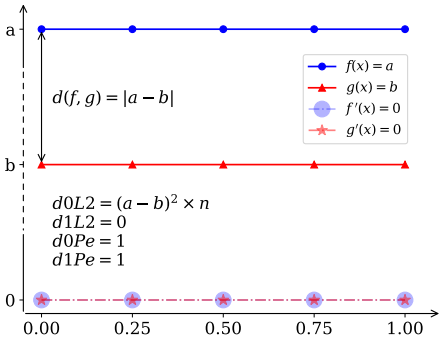
Definition 7.1.1 (**Sum Squared Error (SSE)).** *SSE is defined as the sum of the squared difference between the experimental f and the simulated G data-points.*

$$SSE = \sum_{i=1}^{n} \left(f(x_i) - G(x_i)\right)^2 \tag{7.1}$$

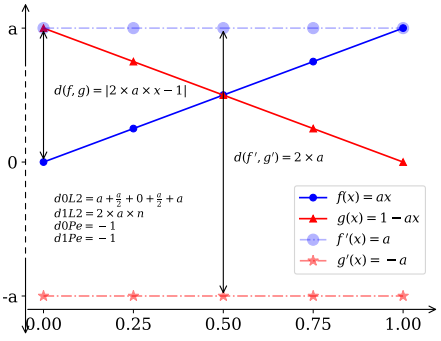
Similar definitions for Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).



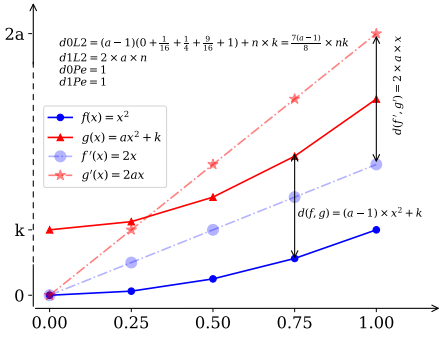
(a) *The pitfall of using point-wise approaches to compute the similarity between the models and the experimental data. This kind of similarity functions could report the same score for the two models even if the trend is different.*



(b) *Curve matching of f and g functions that differ for a vertical translation $|a-b|$. This setting can be understood a posteriori by looking at the curve matching indexes since the $d^0_{L_2} \neq 0$ and $d^1_{L_2} = 0$.*



(c) *Curve matching of two symmetric monotone functions. Except for the sign, they share the same first derivative, in fact $d^0_{Pe} = -1$ and $d^1_{Pe} = -1$.*



(d) *Curve matching of two functions that differ only for a multiplicative factor. In fact $d^0_{Pe} = 1$ and $d^1_{Pe} = 1$.*

Figure 7.2: *Pitfall of the point-wise approaches (Figure 7.2a), and same explanatory examples of curve matching between two functions (Figures 7.2b to 7.2d).*

Curve Matching (CM) is a quantitive *trend approach*, that overcames the limitations of the point-wise approaches, accounting also for the fact that each data point is a part of the trend. CM measures the similarity of two functions f and g with a score $\in [0, 1]$, where 1 is the perfect similarity, after normalization. A detailed description of the CM definition is available in the work by Pelucchi et al. [120].

CM (Definition 7.1.2) is defined as the arithmetic mean of five indexes, $d^0_{L_2}$, $d^1_{L_2}$, d^0_{Pe}, d^1_{Pe}, S. From a modeling point of view, using these indices has different advantages. The Pearson indexes, both on the function d^0_{Pe} and on its first derivative d^1_{Pe} and the SSE computed on both the function $d^0_{L_2}$ and the first derivative $d^1_{L_2}$ capture whether the model trend agrees or disagrees with the experimental data, while the SSE on the function still quantifies the difference point-to-point. The shift S instead measures if the two functions are horizontally translated and are weighted twice since it accounts for both the left and right horizontal shifts. CM also accounts for experimental uncertainty, using a bootstrapping procedure [120]. If the uncertainty is not provided, Curve Matching uses a default uncertainty as suggested in the work of Olm et al. [115].

Definition 7.1.2 (Curve Matching (CM)).

$$CM(f, g) = \frac{d^0_{L_2} + d^1_{L_2} + d^0_{Pe} + d^1_{Pe} + 2S}{6} \tag{7.2}$$

Figures 7.2b to 7.2d show examples curves' comparison using the same indexes used by CM without normalizing neither the values of the indexes or that of the curves. In all the examples, for simplicity, but without losing generality, the axes are adimensional and the x-values range from 0 to 1.

7.1.2 Analysis

The third and last macro phase of the proposed evaluation procedure consists of analyzing the similarity indexes computed during the model validation. The model analysis leverages data science techniques to collect knowledge about the predictive model's behavior systematically.

As in many computer science applications, also in this case, there is a need to address what is known as the "curse of dimensionality" [211]. In fact, for the model validation, each pair of experimental and simulated data is computed as a similarity score, and an average of all of them is a fair indicator of the general model performance. However, such an average cannot provide detailed information about the behavior of the predictive

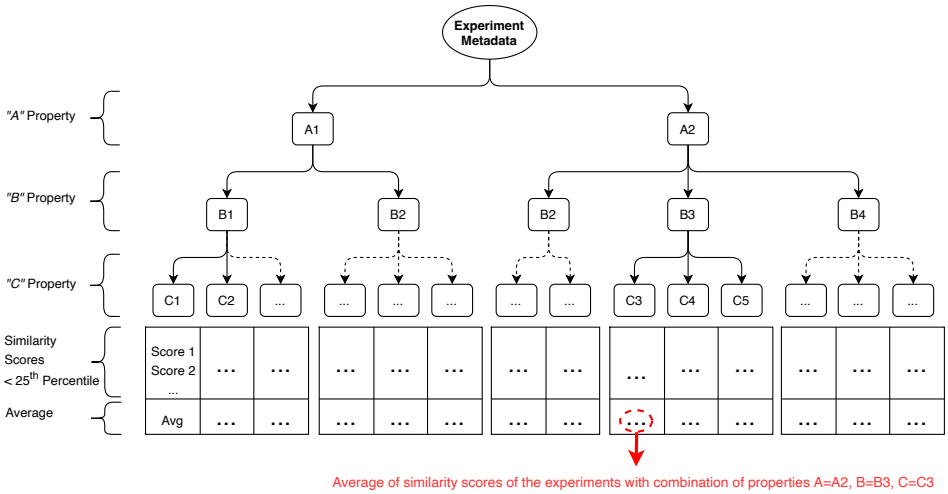


Figure 7.3: *Subdivision and average of the similarity scores of the experiments with the same combination of categorical attributes (A, B, C).*

model since it depends on many variables. Applying data science techniques allows for managing the many dimensions that define a domain and the thousands of similarity indexes computed during validation to extract insights automatically.

The analysis phase is characterized by the following three steps that can be used to study the model simulation results more in-depth: statistical analysis, pattern detection, and interval analysis.

Statistical Analysis Experimental data are provided with additional information (also referred to as characteristics, metadata, or properties) that can be leveraged to statistically analyze the model performances in a complex and multidimensional domain. First, it is possible to group the similarity indexes based on common characteristics of the experiment to know *which* combination of them indicates the worst model performance. Second, using correlation on the experiment metadata, it is possible to investigate *why* the model does not perform well enough, i.e., outside the experimental uncertainties, or in other words, the contributing causes.

First of all, a collection of experiments is filtered based on their similarity score, whether it is below the first quartile (25^{th} percentile or 1^{st} quartile). The percentile is computed with respect to the global distribution of similarity scores. In such a way, the focus is immediately shifted to the experiments with the associated worst predictive model performance.

However, it is necessary to find out *which* combinations of the proper-

ties of the experiments correspond to such behavior. Without losing generality, let us assume that each experiment is characterized by the values assumed in correspondence of three categorical properties A, B, C. Since each property is categorical, only a precise set of values can be assumed, and not all combinations of property values are possible in a domain. Let us assume, for instance, that the values for each category are defined as follows: $A = \{A1, A2\}$ $B = \{B1, B2, B3, B4\}$, $C = \{C1, \ldots, C5\}$.

Following the idea pictured in Figure 7.3, it is possible to compute the average (or other statistical measures) of the similarity scores of the experiment that have a precise combination of experimental properties. Then the combinations of properties that are statistically relevant are observed. A combination is statistically relevant whether it has a considerable number of cases in the quartile and a high percentage of them with respect to the total number of existing ones with that combination of properties.

In a second moment, a correlation analysis between all the experiments metadata, such as type of experiment and environmental conditions, together with the similarity score, suggests, for example, that the model performance is due the use of particular conditions (like equipment) when a specific variable (species for example) is measured. To this purpose, both clustering and classification techniques can be adopted to analyze the results on a large scale.

The arithmetic mean, median, and standard deviation are mainly used as statistical indexes for this work. In addition, the Pearson correlation [212], the point-biserial [213] and the logistic regression [214] are used when it is needed to correlate two variables that could be continuous or categorical. All correlation indexes range from -1 to 1, where 1 indicates two closely and positively correlated variables.

Pattern Detection Pattern detection algorithms, such as clustering, applied to the similarity index, together with (numerical and continuous) physical properties associated with an experiment such as temperature and pressure, can automatically distinguish the portions of the domain *where* the model does show larger mispredictions. Clustering algorithms group similar experiments in the same cluster: taking the most representative cluster(s) with the lowest variation of the associated performance scores, it is possible to know which combination of physical property range is responsible for the worst performance. Data mining is a field of data science that applies a series of techniques to extract hidden features from large quantities of data. In particular, pattern detection or recognition is the process of discovering patterns and regularities in the data. Clustering is a typical unsupervised

machine learning algorithm that allows examining a collection of data and, given a measure of distance, groups them into clusters based on their similarity [211]. Once the data are organized in clusters, it is possible to analyze their common features and understand the pattern, the discriminant that has brought the data together. In this work, Affinity Propagation [215] is used as a hierarchical clustering algorithm. Affinity Propagation selects a number of samples from the dataset as representatives of all the others. The algorithm exchanges a message between pairs of samples to determine which one is suitable to represent the other one. Representatives are continuously selected until convergence, at which point the final clusters are given. Affinity Propagation, by definition, establishes the number of clusters based on the data provided. However, two parameters need to be set: the preference, which controls how many exemplars are used, and the damping factor, which controls the message flow, damping some of them to avoid numerical oscillations.

Interval analysis Once the analysis has been identified for *which* combination(s) of metadata and *where* the model is more deficient, with the developed *ad-hoc* analysis of the intervals, it is possible to quantify (*how much*) the average deviation of the experimental curve from the simulated one.

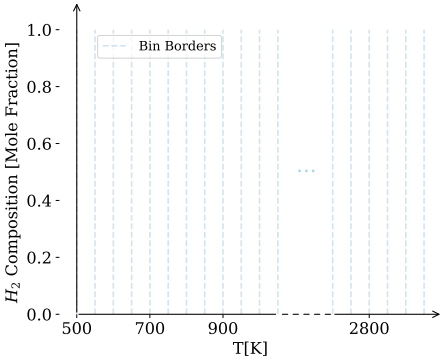
CM, working with a large number of data, provides a synthetic score about how good a model is. However, this synthetic result hides the detailed behavior of a predictive model. Instead, interval analysis, given a set of experimental and simulated data, computes the error of the model in predicting specific targets, in terms of quantitative overestimation and underestimation, in different ranges of a physical property (e.g., temperature). The basic idea of dividing the physical domain into intervals for different purposes has been used several times in the literature. However, either they use a point-wise similarity score to assess the model performance in an interval [114], or, leveraging the concept of data consistency and constraint definition [216, 217], they identify a region in the domain called "feasible set" in which a model can be generated and optimized [218, 219]. Interval analysis, instead, uses a trend similarity score and measures the model performance in each interval for model validation purposes. In other words, curve matching summarizes the similarity between two curves, while the interval analysis maintains the axial dimension and quantifies the overestimation or underestimation of one curve with respect to the other.The disadvantage of maintaining a physical dimension comes with the curse of dimensionality. However, in the procedure proposed in this paper, this algorithm is used as the last step. The previous analyses have identified the

single physical dimension and group of experiments (in terms of common features) that significantly impact the model performance.

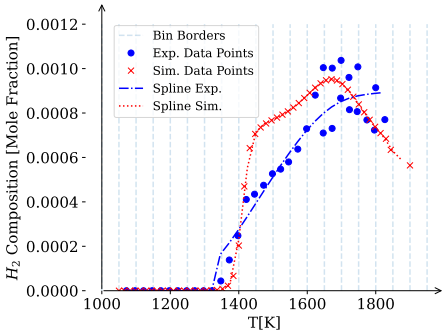
Given as input a set of experimental and simulated data pairs having the same variable on the abscissa and ordinate axes as input, the interval analysis algorithm is divided into four phases. (i) (Figure 7.4a). Given the independent variable operative domain, it is divided into n parts. The division could be equally distributed or not. For example, if the independent variable is the temperature and has an operating domain from 500K to 2500K, this dimension can be divided into 200 sectors or *bins*, each of 10K, such as (500K, 510K), (510K, 520K), and so on. (ii) (Figure 7.4b). For each pair of experimental and simulated data, their corresponding splines are generated. (iii) (Figure 7.4c). For each bin in which the experimental and simulated splines are defined, the area underlying the sub-portion of the domain delimited by the bin's ends is calculated. Then, the ratio of the two areas is calculated and stored, providing a precise quantification of overestimation or underestimation of the simulated data concerning the experimental data. (iv) (Figure 7.4d). Following the previously described procedure, once each pair is analyzed, the model behavior can be summarized by averaging the ratios for each bin, distinguishing for each case whether it is an underestimation or an overestimation. The result of such analysis provides punctual information about the model behavior as the value on the x-axis changes.

7.2 Adaptive Sampling

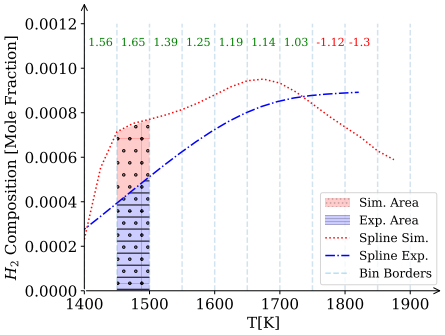
Predictive models are developed following a data-driven black-box or white-box approach. However, regardless of the methodology, the time required for prediction or simulation can be particularly computationally expensive and, thus, time-consuming. The development of an approximation of predictive models can overcome this limitation. Such approximations are called metamodels and are characterized by a shorter prediction time, even by several orders of magnitude. Their development involves three phases: 1) DOE, also called the sampling phase, 2) development of a metamodel, and 3) verification of the metamodel [155]. The DOE aims to select the smallest and most informative set of points from the domain on which the predictive model is evaluated. These data are then the training set on which a metamodel is generated using, for instance, the kriging interpolation method [220]. Finally, the metamodel approximation performances are evaluated against a validation test. If the metamodel approximates the predictive model well, it can be employed in many time-consuming applications, such as in optimization processes, to find the optimal solution.



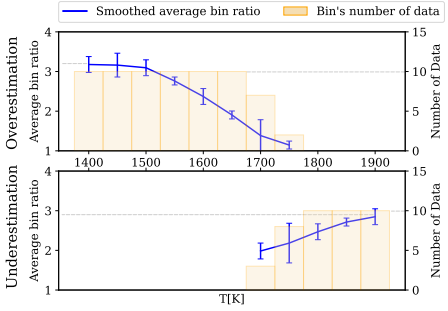
(a) *Step 1. Division of the x-axis dimension (Temperature, in this case) into bins, while the y-axis measures a given property (H_2 composition, in this case)*



(b) *Step 2. Starting from the experimental and simulated data points, the experimental and simulated splines are generated.*



(c) *Step 3. Zoom-in on a portion of the x-axis of the example in Step 2. The ratio between the experimental and simulated area under the curve for each bin is computed. The simulation overestimates (in green) the experimental data if the ratio is bigger than one. Otherwise, it is underestimating (in red). Steps 2 and 3 are repeated for every available pair (in this case, T vs. H_2) in the database.*



(d) *Step 4. All the results of Step 3 are aggregated in order to know, on average, the amount of overestimation or underestimation in each bin. Error bars represent the standard deviation from the mean value. Bar plot represents the number of available pairs in a bin. In this case, the model tends to overestimate by a factor 3 in the low temperature, while at higher temperature underestimates.*

Figure 7.4: *The four steps of the interval analysis procedure.*

This work focuses on the DOE phase, thus on adaptive sampling algorithms. These algorithms iteratively refine the selection of new points to sample and can also be leveraged to optimize the training resources needed to build a good predictive model. Selecting the smallest and most informative training set can reduce the resources needed to collect the data and train a predictive model.

Given a function (or response surface) over a domain (or parametric space) $\mathbb{X}$, $f : \mathbb{X} \rightarrow \mathbb{Y}$, where the input $x \in \mathbb{X} \subset \mathbb{R}^n$ and the output is $y \in \mathbb{Y} \subset \mathbb{R}$. A surrogate model, or metamodel $\mathcal{M}$ approximates f with a loss $\mathcal{L}(f, \mathcal{M}, \mathbb{X})$. Given a point $x \in \mathbb{X}$, $\tilde{y} = \mathcal{M}(x)$ represents the value predicted by the metamodel. Therefore, given a training dataset (or set of samples, or experimental design) of m points $\mathcal{X} = \{x^1, ..., x^m\} \subseteq \mathbb{X}$, the corresponding training data set will be $\mathcal{D} = \{(x^i, y^i), \quad i = 1, ..., m\}$ where $y^i = f(x^i)$. The goal of adaptive sampling is to select and include, at each iteration, following an adaptive strategy, a new $x \in \mathbb{X}$ to $\mathcal{X}$ ($\mathcal{X} \leftarrow \mathcal{X} \cup x$) such that $\mathcal{M}$ better approximate f after training on the new sample dataset $\mathcal{D}$ ($\mathcal{D} \leftarrow \mathcal{D} \cup (x, f(x))$. This procedure is repeated until a termination condition is reached, such as the size of the sampled dataset $\mathcal{D}$ or a loss $\mathcal{L}(f, \mathcal{M}, \mathbb{X})$ value. Kriging is a standard algorithm used in the literature to reconstruct the original response surface into model $\mathcal{M}$ given the training dataset $\mathcal{D}$ selected by the adaptive sampling algorithm [149].

The trade-off in the generation of a surrogate model $\mathcal{M}$ with an adaptive sampling algorithm is to use the smallest set $\mathcal{X}$ that minimizes the loss $\mathcal{L}$. In other words, from a DOE perspective, which is the most informative point x that determines the biggest improvement in terms of $\mathcal{L}$ for $\mathcal{M}$. As explained in Chapter 2, the literature identifies two major goals for the design of an adaptive sampling algorithm: global meta-modeling and optimization. In the first case, the focus is on the metamodel errors and selecting the training set that best estimates the response surface over the entire domain. In the second case, the sample points are selected from particular regions of the response surface domain, optimizing the location of local and global minimum and maximum points. During the design of an effective adaptive sampling algorithm, global exploration, local exploitation, and the trade-off between these two should be considered [151, 221]. With global exploration, the selection of the next point to be sampled reduces the dimension of significant unexplored domain areas by implementing, for instance, distance criteria on the already available sample points. Local exploitation, instead, guides the selection of the next point based on the information available at the moment, such as the approximation error. Designing a DOE algorithm should consider and balance both aspects cor-

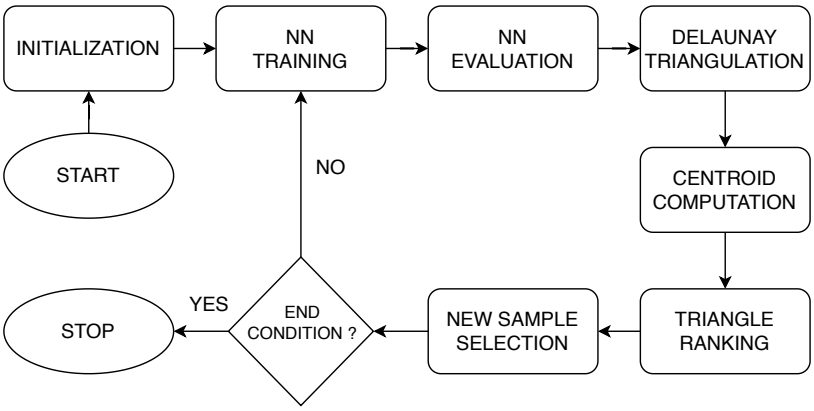


Figure 7.5: *MADO algorithm flowchart.*

rectly. Otherwise, the sampling performance and the metamodel trained on these points will likely approximate the predictive model worse.

This work proposes an adaptive sampling algorithm, Multi Adaptive Delaunay Optimization (MADO), that considers all three aspects, unlike the other adaptive sampling algorithms presented in Chapter 2. It combines computational geometry, machine learning techniques, and a strategy to combine these two. The generalization capabilities of NNs are leveraged for the exploitation component. Instead, for the exploration part, the Delaunay triangulation conveys information about the unexplored areas with the biggest representation errors. The results show competitive performance, whether the goal is global metamodeling or optimization. The main steps of the algorithm are shown in Figure 7.5 and presented in the following. Figure 7.6 shows a visualization of the algorithm steps at a given iteration.

Delaunay triangulation is a space partitioning algorithm. Given a set of $\mathcal{P} = x_1, ..., x_n, \quad x_i \in \mathbb{R}^2$ distinct ($|\mathcal{P}| = n \geq 3$) points in a two-dimensional space, the Delaunay triangulation of $\mathcal{P}$, $\mathcal{T}(\mathcal{P})$, triangulates the points $\mathcal{P}$ such that no point in $\mathcal{P}$ is inside the circumscribed circle of any triangle in $\mathcal{T}(\mathcal{P})$. Given h as the number of convex hull of $\mathcal{P}$, the Delaunay Triangulation $\mathcal{T}(\mathcal{P}) = \{\triangle_1, ..., \triangle_k\}$, generates $k = 2n - h - 2$ triangles $\triangle$, where each triangle is a triple of vertices $\mathcal{V} \in \mathcal{P}$, and $\triangle_i = (x_l, x_m, x_j), \quad x_l, x_m, x_j \in \mathcal{P}$ [222].

Initialization The MADO algorithm requires only a few starting points in $\mathcal{X}$ located at the corners of the parametric space (or domain) $\mathbb{X}$ of the function f to be sampled, together with the corresponding $y-$values, i.e., the sample set $\mathcal{D}$. Without points $\mathcal{X}$ on the domain boundaries, the Delaunay

triangulation $\mathcal{T}(\mathcal{X})$ would not cover the entire parameter space, potentially excluding important regions. Without losing generability, this initialization requirement can be combined with other typical techniques, such as Latin Hypercube Design (LHD) (or also known as Latin Hypercube Sampling (LHS)) or random sampling.

Neural Network Training A NN $\mathcal{N}$ is trained to approximate the function of interest f, using as the training set $\mathcal{D}$, i.e., the available sampled points as expressed in Equation (7.3).

$$\mathcal{N} \leftarrow \text{Train}(\mathcal{D}) \tag{7.3}$$

The complexity of the NN, in terms of its architecture and hyperparameters, can significantly impact the algorithm's effectiveness. A more complex network or a more prolonged training procedure may be able to approximate a complex function better but at the cost of increased computational time and the risk of overfitting. Conversely, a simpler network may be faster and less prone to overfitting, but it might not capture complex function behavior. A good trade-off between architecture complexity and appropriate training procedure can generate an NN good representation of the domain while highlighting the complex domain area to represent.

Neural Network Evaluation The NN acts as an approximation function that generalizes based on the limited information provided. When the true result deviates significantly from the NN prediction, it suggests an area of the function domain that is more challenging to approximate. These areas, highlighted by larger errors from the NN, are targeted for denser sampling in the next iteration. This strategy efficiently directs the algorithm's attention towards "difficult" areas, contributing to the exploitation part of the algorithm. Equation (7.4) reports the computation of the error of the NN $\mathcal{N}$ and the function f over $\mathcal{X}$.

$$\mathcal{E}(i) = |f(x_i) - \mathcal{N}(x_i)| \quad x_i \in \mathcal{X}, i = 1, ..., |\mathcal{X}| \tag{7.4}$$

Delaunay Triangulation Once the NN is trained, MADO constructs a Delaunay triangulation $\mathcal{T}(\mathcal{X})$ of the current set of sampled points $\mathcal{X}$. Each triangle $\triangle = \{\triangle_1, ...\triangle_k\} \in \mathcal{T}(\mathcal{X})$ represents a region of the parameter space. k is the number of Delaunay triangulations. $\mathcal{A}_i$ is the area (or volume if in the n-dimensional space) of each triangle $\triangle_i$. The area of the triangle is the explorative component of the algorithm.

Centroid Computation For each triangle $\triangle_i = (x_l, x_m, x_j) = (V_{i,1}, V_{i,2}, V_{i,3})$, $x_l, x_m, x_j \in \mathcal{X}$ in the Delaunay triangulation $\mathcal{T}(\mathcal{X})$, this steps computes the centroid C_i as a weighted mean of the vertices, where the weights are the absolute errors at the vertices as shown in Equation (7.5). The centroids tend to move towards the vertices with the highest prediction error, thus directing the algorithm towards regions where the function approximation can be most improved.

$$C_i = \frac{\sum_{j=1}^{3} \mathcal{E}(V_{ij}) V_{ij}}{\sum_{j=1}^{3} \mathcal{E}(V_{ij})} \tag{7.5}$$

For each $\triangle_i$, $\hat{\mathcal{E}}_i$ is the total error of the triangle computes as presented in Equation (7.6).

$$\hat{\mathcal{E}}_i = \sum_{j=1}^{3} |\mathcal{E}(V_{ij})| \tag{7.6}$$

Triangle Ranking The algorithm then computes a score $\mathcal{S}_i$ for each triangle $\triangle_i$ that represents the priority to sample a new point from it. This score is a function of both the triangle's area $\mathcal{A}_i$, which represents the size of the region it covers and the total estimated prediction error of the neural network within this triangle $\hat{\mathcal{E}}_i$. This dual criterion aims to balance with the parameter β the need to explore large unsampled domain regions and exploit areas where the current function approximation is more uncertain and, thus, more challenging to represent.

$$S_i = \frac{(\mathcal{A}_i^{1-\beta} \hat{\mathcal{E}}_i{}^{\beta})}{\sum_{j=1}^{k} \mathcal{A}_j^{1-\beta} \hat{\mathcal{E}}_j{}^{\beta}} \tag{7.7}$$

New Sample Selection The last step of the algorithm loop concerns the selection of the new point(s) $x \in \mathbb{X}$ to be sampled and added in $\mathcal{X}$ and then in $\mathcal{D}$ after computing $f(x)$. According to Equation (7.8), the next point to be sampled will be from the triangle $\triangle_i$ with the highest score $\hat{\mathcal{S}}$. The value of the new point $x \in \mathcal{X}$ will be equal to the centroid $\mathcal{C}_i$ computed previously in Section 7.2, following Equation (7.5).

The number of points selected in each iteration can significantly impact the efficiency and effectiveness of the algorithm. If only one point is selected, the algorithm will need more iterations to sample the function's

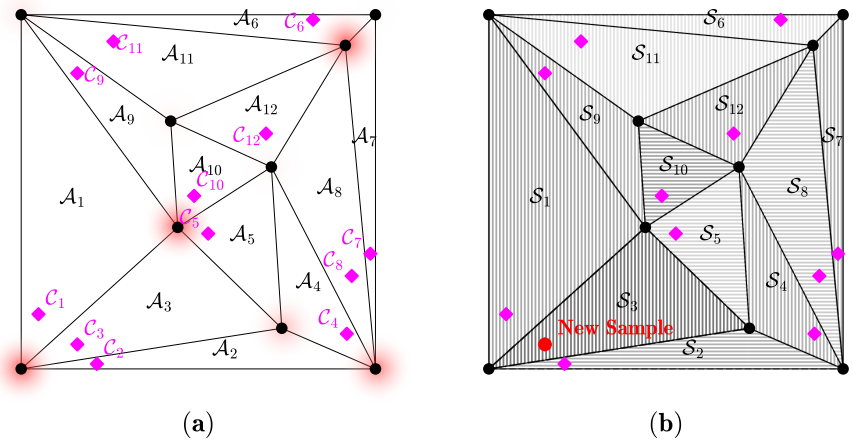


Figure 7.6: *Visual representation of the MADO algorithm steps at a certain iteration. (**a**) It describes the NN evaluation on the already available sample points, the Delaunay triangulation, and the centroid computation. (**b**) It depicts triangle ranking and the selection of the next sample point.*

domain, increasing computational load and execution time. However, selecting one point per iteration allows it to adjust its adaptive strategy more frequently based on the new information, potentially leading to a more refined exploration and exploitation process. On the other hand, if the stop criterion is to achieve a given number of points in $\mathcal{D}$, selecting more points in each iteration reduces the total number of iterations needed, thus potentially reducing the total computational time. Moreover, selecting multiple points can lead to a broader exploration of the function's domain in each iteration, as the selected points are likely to cover a wider range of the domain. As a drawback, this approach might also lead to less exploitation behavior. Therefore, selecting one or more points in each iteration represents a trade-off between computational efficiency, exploration, and exploitation.

$$\hat{S} = \arg\max_{i} \mathcal{S}_i \tag{7.8}$$

Following the literature practices [149], the validation methodology is as follows. The evaluation metrics include the Normalized Root Mean Squared Error (NRMSE), Normalized Mean Absolute Error (NMAE), Minimum Normalized Mean Absolute Error (NMAE_{min}), and the coefficient of determination (R^2). This thesis employs the Gaussian Process Regressor (GPR) [220] to approximate the target function for all experiments. The MADO algorithm is tested against 14 test functions to generate 100 points

as training data across the function domain and is evaluated over a 200×200 grid of the domain used as validation dataset. A visualization of the sampled points by the MADO algorithm and the original and the reconstructed function is depicted in Figure 7.7.

The experimental design comprises two objectives: assessing reconstruction quality (Section 7.2.1) and analyzing algorithm features (Section 7.2.2). For the first objective, this thesis conducts a comparative analysis to assess the performance of the proposed adaptive sampling algorithm against conventional sampling methods [223], such as LHD with both basic [224] and maximin sampling ($\text{LHD}_{maximin}$) criteria [225, 226]. Thus, it evaluates the reconstruction quality achieved by GPR models trained on datasets generated through these three sampling methods. This analysis enables to determine the ability of each algorithm to select informative samples for reconstructing accurate surrogate models. The second goal is to investigate the behavior of MADO in balancing the exploration or exploitation component by adopting different criteria for calculating the β parameter according to the *Ramp*, *Exponential*, or *Sigmoid* function, normalized between 0 and 1, and changing the number of selected points per iteration (1, 5, and 10). The latter characteristic determines the granularity and speed at which the algorithm explores the domain space. For instance, by selecting more points at each iteration, the algorithm could reach the termination condition quicker but potentially at the expense of worse accuracy in specific space areas. Both characteristics can be tuned to improve the algorithm's performance across different applications and scenarios.

7.2.1 Reconstruction Quality

Table 7.2 and Table 7.3 present a detailed illustration of each technique's performance metrics across the 14 test functions. From these tables, several insights into the comparative performance of the LHS (both basic and maximin variations) and the proposed MADO sampling method can be drawn. Firstly, in the majority of test functions, MADO consistently outperforms both versions of LHS in terms of NMAE, NMAE_{min}, NRMSE, and R^2, suggesting that MADO is more proficient in capturing the underlying behavior and complexities of the functions under examination. The particularly lower values of NMAE and NRMSE for MADO in functions such as *DropWave*, *Franke* (Figure 7.8), and *Griewank* emphasize an ability to accurately represent areas where the functions have sharp variations, peaks, troughs, or changes in the sign of their derivatives. Furthermore, the R^2 values underline MADO's efficiency, as it consistently records values closer

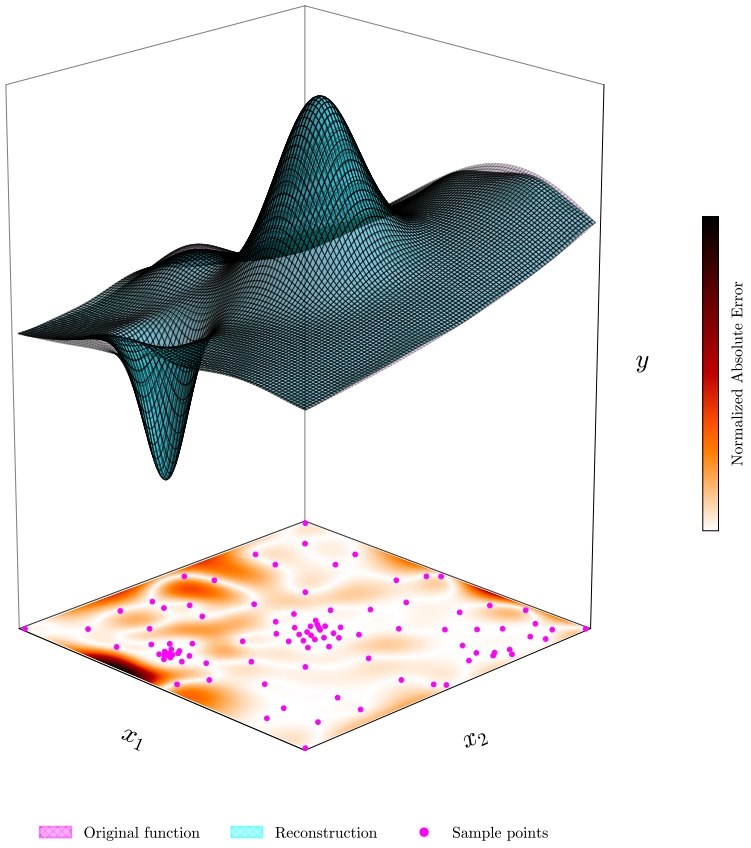


Figure 7.7: *A comparison between the original test function (magenta surface) and its reconstruction achieved by the MADO algorithm (cyan surface). The magenta dots represent the sampled points used for reconstruction. The background displays the distribution of normalized absolute error between the original and reconstructed surfaces, with darker regions indicating higher error. The plot highlights MADO's effectiveness in generating informative sample points for accurate surrogate model construction.*

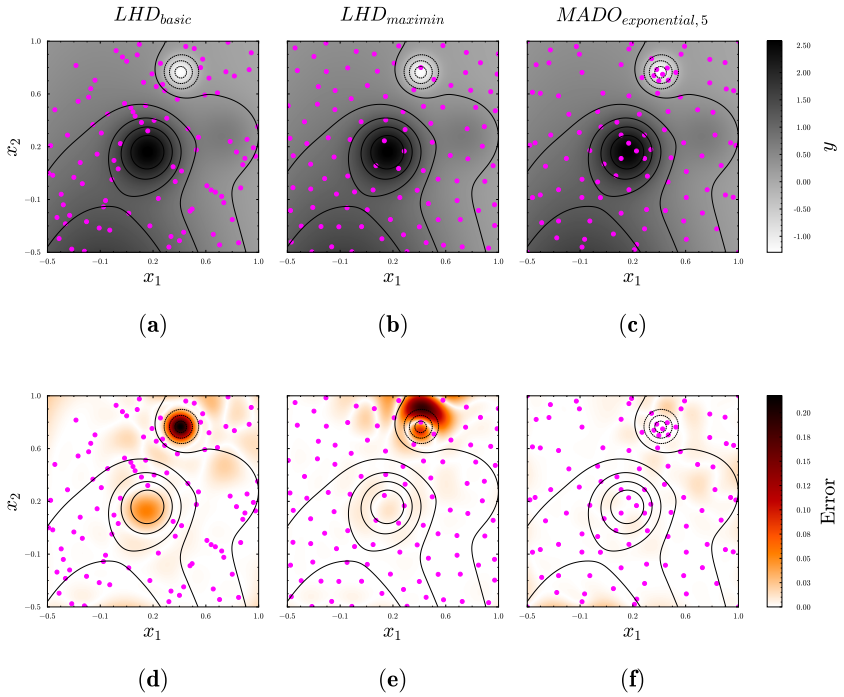


Figure 7.8: *Visualization of sample placements for the Franke function across three sampling methods.* ***(a)*** *The LHD exhibits uniform distribution across intervals.* ***(b)*** *$LHD_{maximin}$ emphasizes a dispersed placement, enhancing point separations.* ***(c)*** *MADO's samples focus on intricate function regions — peaks, troughs, and sign-changing derivative points, indicating its precision.* ***(d)*** *LHD reveals its limitations in normalized absolute error assessment against the Franke function, especially in complex regions.* ***(e)*** *$LHD_{maximin}$, despite its spread, presents errors in nuanced function regions.* ***(f)*** *MADO showcases reduced and uniformly distributed errors, emphasizing its detailed sampling capability.*

	NMAE			$NMAE_{min}$		
	LHD_{basic}	$LHD_{maximin}$	MADO	LHD_{basic}	$LHD_{maximin}$	MADO
Ackley	0.009	0.006	**0.003**	0.272	0.716	**0.021**
Bird	0.049	0.039	**0.026**	0.207	0.066	**0.001**
Bohachevsky	0.011	**0.006***	**0.006***	**0.000**	0.001	0.004
Booth	0.004	**0.002***	**0.002***	**0.001***	**0.001***	**0.001***
Branin	0.033	0.015	**0.009**	0.006	0.007	**0.000**
DropWave	0.152	0.272	**0.004**	0.013	0.341	**0.001**
Franke	0.018	0.010	**0.002**	0.058	0.348	**0.000**
Griewank	0.218	0.215	**0.076**	0.334	0.288	**0.036**
Himmelblau	0.045	**0.014**	0.019	0.009	0.004	**0.001**
Levy	0.108	0.123	**0.091**	0.016	0.018	**0.016**
Michalewicz	0.082	0.081	**0.047**	0.425	0.445	**0.341**
Rastrigin	0.101	0.105	**0.097**	0.127	0.138	**0.000**
Rosenbrock	0.028	0.015	**0.011**	0.030	0.067	**0.024**
SixHumpCamel	0.175	0.074	**0.009**	0.042	0.260	**0.004**

Table 7.2: *Comparison of MADO (our) with LHD_{basic} and $LHD_{maximin}$ according to the NMAE and $NMAE_{min}$ metrics against 14 test functions. Values in bold represent the best score achieved for each respective test function (row), while asterisks (*) denote the overall best score across all rows. Please note that due to rounding, some values may appear identical, but only one holds the best score.*

to 1, suggesting better prediction capabilities than the other two methods. This distinction is more evident in *Ackley*, *DropWave*, and *SixHumpCamel* functions.

Considering the LHS variants, the difference between LHS_{basic} and $LHS_{maximin}$ shows varied results across functions. However, $LHS_{maximin}$ is generally slightly better, most likely due to its enhanced dispersion of samples, as in the cases of *Bohachevsky* and *Himmelblau*.

7.2.2 MADO's features

Tables 7.4 to 7.6 and section 7.2.2 show the performance result obtained by varying both the β calculation criterion and the number of points selected in each iteration. Instead, Figure 7.9 directly compares the different β calculation criterion with the same number of sampled points for each algorithm iteration for the *DropWave* test function. The Delaunay triangulation, central to MADO, inherently imposes a spatial structure on the sampling. Adding multiple points within a single iteration ensures that these points are sampled in separate triangles, hence distinct regions of the parametric space. This mechanism disperses the sample points, leading to a geometrically spread-out distribution, thereby bolstering the explorative characteristic of the algorithm. On the other hand, opting for a single point per iteration allows the algorithm to delve deeper into a specific region by splitting existing triangles. Moreover, when sampling more points at each

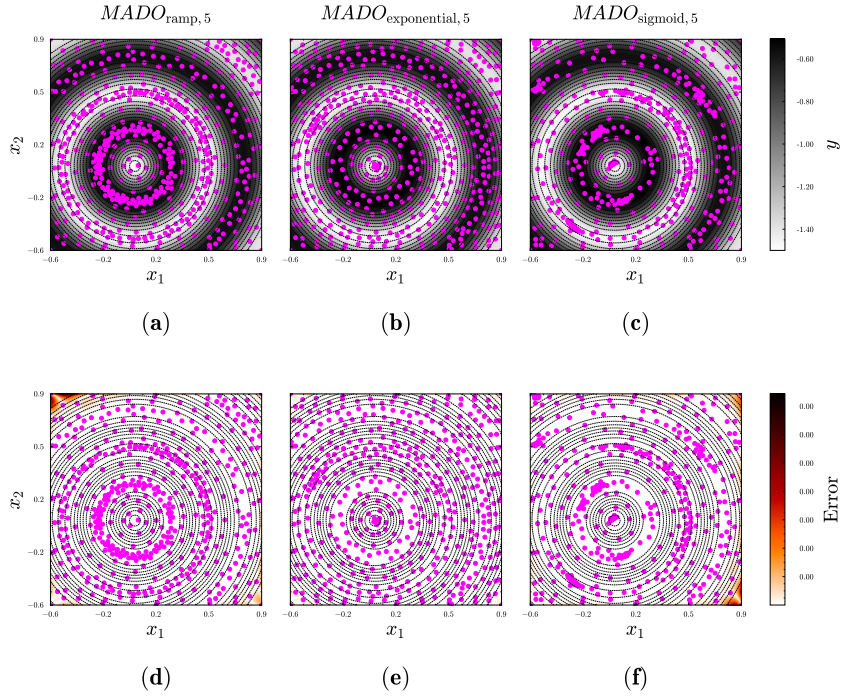


Figure 7.9: *Sampling strategies on the DropWave function and associated kriging model errors.* ***(a-c)*** *Display 500 samples within* $[-0.6, 0.9] \times [-0.6, 0.9]$ *using Ramp, Exponential, and Sigmoid criteria.* ***(a)*** *(Ramp) densely targets the function's extremes, occasionally sidelining peripheral regions.* ***(b)*** *(Exponential) strikes a harmonious balance between exploration and exploitation, ensuring a more equitable point distribution.* ***(c)*** *(Sigmoid) accelerates from exploration to intense exploitation, closely resembling the Ramp's concentration on function's critical points.* ***(d-f)*** *Depict the reconstruction errors from fitting a kriging model to these samples.* ***(d)*** *and* ***(f)****, corresponding to Ramp and Sigmoid, achieve high precision in primary areas but show discrepancies towards the domain's edges. Conversely,* ***(e)*** *(Exponential) maintains a uniformly distributed error, reflecting its balanced sampling approach.*

	NRMSE			R^2		
	LHD_{basic}	$\text{LHD}_{maximin}$	MADO	LHD_{basic}	$\text{LHD}_{maximin}$	MADO
Ackley	0.041	0.032	**0.005**	0.356	0.601	**0.990**
Bird	0.074	0.056	**0.042**	0.766	0.868	**0.926**
Bohachevsky	0.021	0.015	**0.012**	0.992	0.996	**0.997**
Booth	0.010	0.004	**0.003**	0.997	0.999	**1.000**
Branin	0.068	0.032	**0.021**	0.820	0.959	**0.983**
DropWave	0.200	0.305	**0.008**	0.618	0.108	**0.999**
Franke	0.047	0.035	**0.003**	0.879	0.934	**0.999**
Griewank	0.263	0.259	**0.134**	0.098	0.126	**0.765**
Himmelblau	0.077	**0.023**	0.031	0.694	**0.972**	0.951
Levy	0.141	0.154	**0.124**	0.407	0.290	**0.545**
Michalewicz	0.122	0.127	**0.076**	0.533	0.494	**0.818**
Rastrigin	0.132	0.133	**0.125**	0.364	0.350	**0.430**
Rosenbrock	0.052	0.029	**0.021**	0.883	0.963	**0.980**
SixHumpCamel	0.339	0.138	**0.018**	0.092	0.709	**0.987**

Table 7.3: *Comparison of MADO (our) with LHD_{basic} and $\text{LHD}_{maximin}$ according to the NRMSE and R^2 metrics against 14 test functions. Values in bold represent the best score achieved for each respective test function (row), while asterisks (*) denote the overall best score across all rows. Please note that due to rounding, some values may appear identical, but only one holds the best score.*

iteration is combined with the *Exponential* beta criterion, which retains a low beta value for most of the iterations, it favors a pervasive exploration of the function's domain. However, this can result in a later placement of the sample points in the maximum or minimum of the test function. The *Ramp* criterion offers an adaptable progression. It initiates with a phase reminiscent of the exponential's broad exploration but can be tailored to shift gears toward exploitation depending on the point where the ramp ascent is set. The *Sigmoid* criterion selects high beta values earlier, resulting in a prolonged exploitation phase than the *Ramp*. This characteristic, especially when paired with multiple-point selections, can expedite the exploitation algorithm's behavior, potentially at the expense of global exploration. In summary, the *Sigmoid*, with its extended high beta phase, might be the frontrunner for optimization-focused endeavors, whereas the adaptable nature of the *Ramp* can serve as an intermediary, fusing the traits of both *Exponential* and *Sigmoid*. The *Exponential* criterion, particularly when paired with multiple point selection per iteration, achieves detailed exploration, which is necessary for global metamodeling. A visualization of these properties is represented in Figure 7.10. However, the ideal configuration remains intricately tethered to the specifics of the target function and the user goal.

This section proposes a new adaptive sampling algorithm that selects the smallest and most informative data set to be included in the training. Such capabilities reduce the computational resources needed to generate

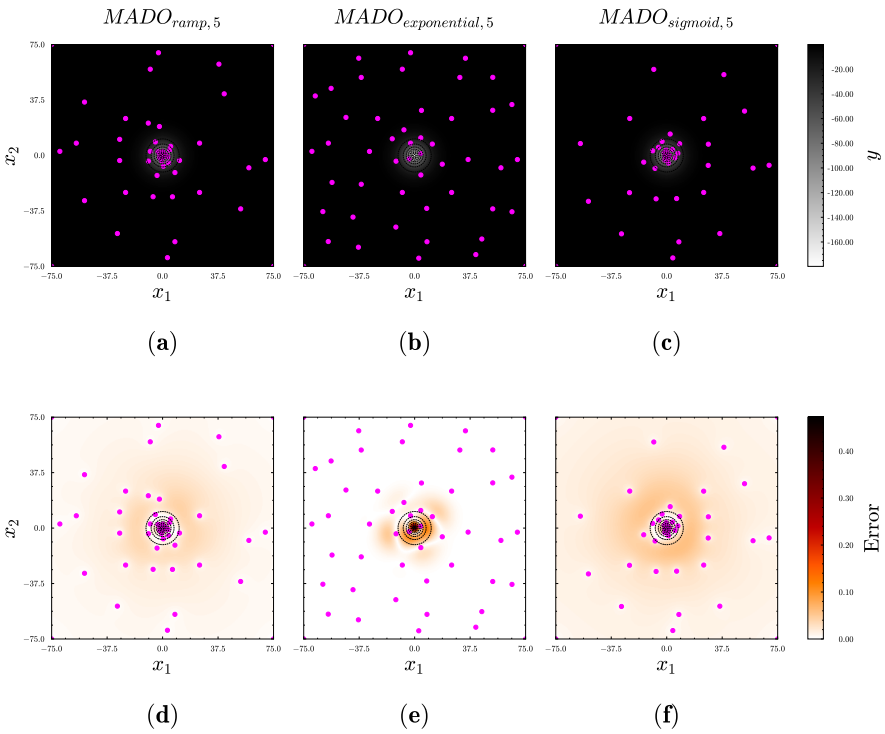


Figure 7.10: *The figure illustrates the sampling results on the Ackley function, characterized by a smooth landscape with a prominent global minimum at the domain's center, using three different sampling criteria. In panels* ***(a)*** *to* ***(c)****, the sampling distributions for the Ramp, Exponential, and Sigmoid criteria are depicted, respectively. The Ramp and Sigmoid criteria both tend to sample near the global minimum. Instead, the Exponential criterion delivers a more uniformly distributed sample set, lacking attention near the minimum. Panels* ***(d)*** *to* ***(f)*** *present the associated reconstruction errors for each criterion. While the Ramp and Sigmoid criteria achieve good precision around the minimum, the Exponential approach, with its sustained explorative strategy, incurs a larger reconstruction error.*

	Exponential			Ramp			Sigmoid		
N_selected	1	5	10	1	5	10	1	5	10
Ackley	0.005	0.003	**0.001**	0.008	0.021	0.016	0.002	0.026	0.019
Bird	**0.026***	**0.026**	0.027	0.032	0.036	0.029	0.037	0.033	0.035
Bohachevsky	**0.006**	**0.006**	**0.006***	0.009	0.019	0.017	0.017	0.033	0.023
Booth	**0.002**	**0.002***	**0.002**	0.005	0.005	0.007	0.005	0.009	0.013
Branin	0.012	**0.009**	0.010	0.016	0.014	0.023	0.022	0.021	0.030
DropWave	0.005	**0.004**	**0.004***	0.010	0.007	0.008	0.024	0.016	0.016
Franke	0.003	**0.002**	0.006	0.003	0.005	0.003	0.006	0.004	0.030
Griewank	**0.064**	0.076	0.066	0.073	0.108	0.090	0.099	0.117	0.141
Himmelblau	**0.019***	**0.019**	0.021	0.022	0.034	0.044	0.034	0.043	0.052
Levy	0.100	0.091	0.091	0.089	0.087	0.088	0.095	**0.086**	0.088
Michalewicz	0.049	**0.047**	0.051	0.067	0.054	0.055	0.072	0.059	0.064
Rastrigin	0.089	**0.086***	0.087	0.095	0.089	**0.086**	0.091	0.093	0.087
Rosenbrock	**0.011**	**0.011***	0.012	0.015	0.019	0.020	0.022	0.025	0.027
SixHumpCamel	0.011	**0.009**	**0.009***	0.012	0.013	0.015	0.014	0.025	0.021

Table 7.4: *NMAE metric scores for the examined test functions. Values in bold represent the lowest NMAE score achieved for each respective test function (row), while asterisks (*) indicate the overall best score across all rows. Please note that due to rounding, some values may appear identical, but only one truly holds the best (lowest) NMAE score.*

or collect training data, addressing the problem of low availability of data in scientific fields and the greediness of machine learning algorithms. The following section will focus on using training data to generate scientific predictive models that are interpretable and do not violate the laws of physics.

7.3 Chemical Reaction Neural Network

This section presents a preliminary study regarding the application of CRNN to develop a predictive model for hydrogen experiments in the field of chemical kinetics. CRNN [227] is a particular NODE [228], that is aimed at learning the reaction pathways combining the generalization capabilities of NN, ingesting data and discovering patterns while ensuring that the fundamental physic laws are fulfilled. CRNN is fully interpretable since the model parameters correspond to the weights learned during the training procedure. Other development of CRNN, B-CRNN [168], also accounts for the experimental uncertainty.

$$\nu_A A + \nu_B B \rightarrow \nu_C C + \nu_D D \tag{7.9}$$

A general elementary reaction is presented in Equation (7.9). Without losing generality, it involves four species $S = \{A, B, C, D\}$ with the corresponding stoichiometric coefficients $\nu_A, \nu_B, \nu_C, \nu_D$.

	Exponential			Ramp			Sigmoid		
N_selected	1	5	10	1	5	10	1	5	10
Ackley	0.864	0.021	0.130	0.029	**0.002***	**0.002**	0.074	0.008	0.005
Bird	0.002	0.001	0.012	**0.000***	0.008	**0.000**	**0.000**	**0.000**	**0.000**
Bohachevsky	0.004	0.004	**0.001***	**0.001**	0.010	0.011	0.008	0.006	0.012
Booth	**0.000**	0.001	0.002	0.001	0.001	0.002	**0.000***	0.001	0.006
Branin	**0.000**	**0.000***	0.001	**0.000**	**0.000**	0.001	0.003	0.002	0.003
DropWave	0.002	**0.001**	**0.001***	**0.001**	0.033	0.005	0.119	0.004	0.027
Franke	0.003	**0.000**	0.204	**0.000**	**0.000**	**0.000**	0.003	**0.000***	0.263
Griewank	0.016	0.036	0.025	**0.003**	**0.003***	0.068	0.019	0.045	0.291
Himmelblau	**0.000***	0.001	**0.000**	0.002	0.001	0.008	0.001	0.003	0.007
Levy	0.019	**0.016**	0.020	0.020	0.020	0.019	0.024	**0.016***	**0.016**
Michalewicz	0.002	0.341	0.010	0.328	**0.000**	**0.000**	**0.000**	**0.000***	**0.000**
Rastrigin	0.078	0.035	0.057	0.036	**0.019**	0.035	0.078	0.031	0.032
Rosenbrock	0.024	0.024	0.029	**0.001**	0.029	0.031	0.032	0.032	0.032
SixHumpCamel	0.003	0.004	0.001	0.008	0.003	0.001	0.003	0.012	**0.000**

Table 7.5: *$NMAE_{min}$ metric scores for the examined test functions. Values in bold represent the lowest $NMAE_{min}$ score achieved for each respective test function (row), while asterisks (*) indicate the overall best score across all rows. Please note that due to rounding, some values may appear identical, but only one truly holds the best (lowest) $NMAE_{min}$ score.*

The law of mass action defines the reaction rate r of Equation (7.9) as in Equation (7.10).

$$r = k[A]^{\nu_A}[B]^{\nu_B}[C]^{\nu_C}[D]^{\nu_D} \tag{7.10}$$

Equation (7.10) can be rewritten as in Equation (7.11)

$$r = exp(Ln\ k + \nu_A Ln\ [A] + \nu_B Ln\ [B] + \nu_C Ln\ [C] + \nu_D Ln\ [D]) \tag{7.11}$$

The formulation of the law of mass action presented in Equation (7.11) has the same structure as the formula of NN $y = \sigma(w\ x + b)$ (Figure 7.11), where the weights w are the stoichiometric coefficients, the bias is the logarithm of the kinetic constant k, and the input x are the concentration of the species $[A], [B], [C], [D]$. The output y corresponds to the formation rates of the concentrations $\frac{d[A]}{dt}, \frac{d[B]}{dt}, \frac{d[C]}{dt}, \frac{d[D]}{dt}$. An ODE solver is able to compute the concentration $[A], [B], [C], [D]$, given their formation rate, thus providing the predicted concentrations by CRNN. With the predicted concentration, it is possible to compute a loss function between the predictions and the real concentrations. The loss function can then be employed in a gradient descent algorithm to optimize the learned value of the stochiometric coefficients, i.e., the network weights w.

If the rate constants are also temperature dependent, it is possible to include in a similar way the Arrhenius law as following Equation (7.12).

	Exponential			Ramp			Sigmoid		
N° selected	1	5	10	1	5	10	1	5	10
Ackley	0.040	0.005	**0.004**	0.010	0.025	0.019	0.005	0.030	0.022
Bird	**0.038**	0.042	0.041	0.052	0.055	0.044	0.060	0.050	0.052
Bohachevsky	0.013	**0.012**	**0.012***	0.016	0.024	0.022	0.029	0.042	0.029
Booth	0.004	**0.003**	0.004	0.008	0.007	0.009	0.008	0.013	0.017
Branin	0.025	0.021	**0.019**	0.026	0.026	0.033	0.036	0.033	0.038
DropWave	0.013	**0.008**	0.009	0.037	0.017	0.021	0.058	0.030	0.034
Franke	0.004	**0.003**	0.020	0.005	0.010	0.005	0.015	0.007	0.069
Griewank	**0.109**	0.134	0.114	0.112	0.177	0.149	0.159	0.186	0.221
Himmelblau	**0.030**	0.031	0.031	0.032	0.043	0.053	0.043	0.052	0.062
Levy	0.132	0.124	0.123	0.115	0.114	**0.113**	0.122	0.115	0.118
Michalewicz	0.081	**0.076**	0.082	0.117	0.090	0.085	0.123	0.091	0.094
Rastrigin	0.111	**0.107**	0.109	0.120	0.111	0.108	0.113	0.116	0.108
Rosenbrock	0.023	**0.021**	0.023	0.026	0.028	0.028	0.035	0.034	0.036
SixHumpCamel	0.021	0.018	**0.017**	0.021	0.023	0.028	0.021	0.038	0.039

Table 7.6: *NRMSE metric scores for the examined test functions. Values in bold represent the lowest NRMSE score achieved for each respective test function (row), while asterisks (*) indicate the overall best score across all rows. Please note that due to rounding, some values may appear identical, but only one truly holds the lowest NRMSE score.*

$$Ln\ k = Ln\ A + b\ Ln\ T - \frac{E_a}{RT} \tag{7.12}$$

This thesis investigates the characteristics of CRNN in different conditions. Given a synthethic reference model in Table 7.8 of four reactions R and five species S. The analysis dimensions regard the quantity of experimental data used for the training, the diversity, and the noise level (or uncertainty) in the experiments. In this preliminary evaluation are performed 392 tests. For each test, it is generated a different training dataset using the synthetic model presented in Table 7.8. The 392 tests are generated according to the following elements: the quantity of training data varies from 20 up to 1500, the wideness of the initial condition on which the experiments are generated varies from 1 to 15, and the noise in the data from 0% up to 40%. An example of noisy experimental data and the corresponding ground truth included in the training set is presented in Figure 7.12.

Figure 7.13 presents the correlation matrix between the analysis dimensions of the test cases, computed using the Kendall correlation. The correlation matrix suggests that CRNN is robust to noisy data and does not require much training data. However, running the test cases, it has been observed that some training instances diverge, as shown in the plot of the loss function in Figure 7.14.

After investigating the source of such problem as in Figure 7.15, the dif-

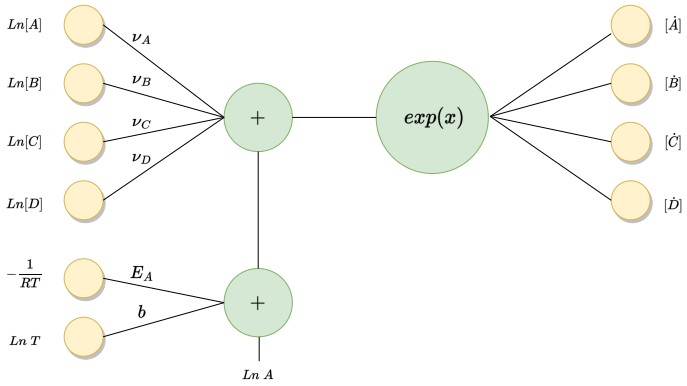
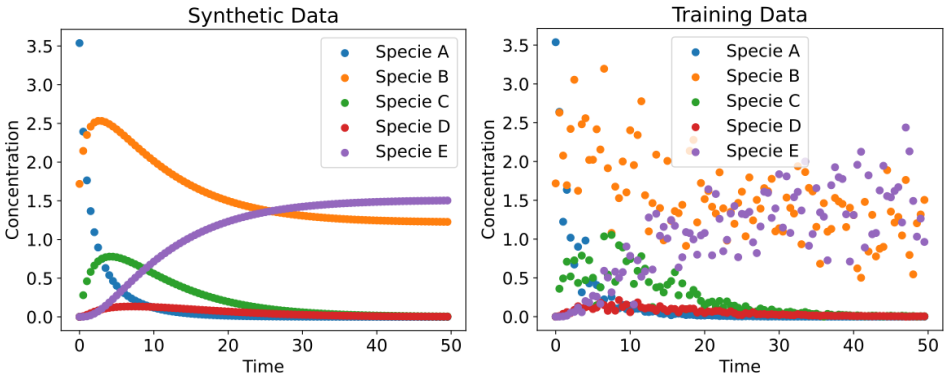


Figure 7.11: *Chemical Reaction Neural Network architecture.*



(a) **(b)**

Figure 7.12: *(a) Example of ground truth of experimental data. (b) Example of experimental data in the training set of a given test case.*

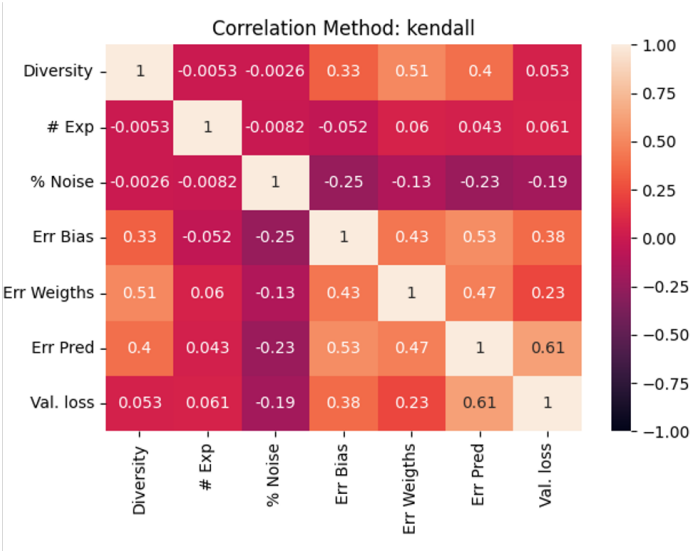


Figure 7.13: *Correlation matrix.*

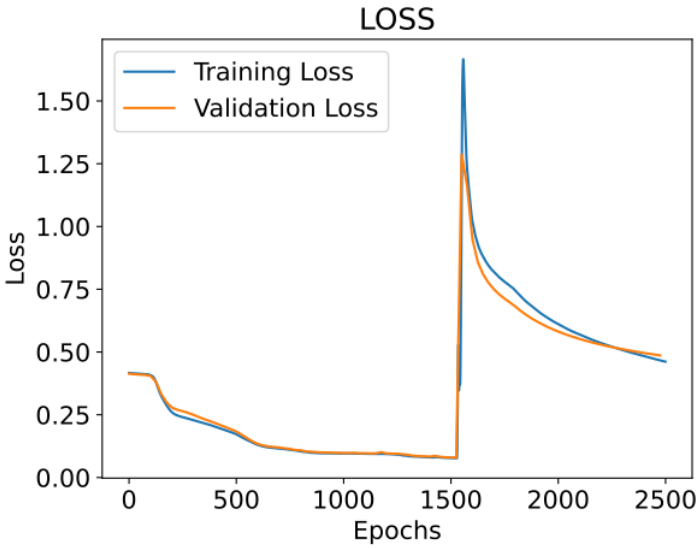


Figure 7.14: *Plot of the loss function for a divergent training instance.*

	Exponential			Ramp			Sigmoid		
N_selected	1	5	10	1	5	10	1	5	10
Ackley	0.377	0.990	**0.994**	0.960	0.753	0.864	0.989	0.650	0.811
Bird	**0.939**	0.926	0.928	0.886	0.873	0.919	0.849	0.895	0.884
Bohachevsky	**0.997**	**0.997**	**0.997***	0.995	0.988	0.990	0.984	0.966	0.984
Booth	0.999	**1.000**	0.999	0.998	0.998	0.997	0.998	0.995	0.991
Branin	0.975	0.983	**0.986**	0.973	0.974	0.958	0.950	0.957	0.943
DropWave	0.998	**0.999***	**0.999**	0.987	0.997	0.996	0.968	0.991	0.989
Franke	**0.999**	**0.999***	0.979	**0.999**	0.994	0.998	0.988	0.997	0.744
Griewank	**0.846**	0.765	0.830	0.836	0.594	0.712	0.672	0.550	0.365
Himmelblau	**0.953**	0.951	0.950	0.946	0.906	0.853	0.904	0.858	0.801
Levy	0.484	0.545	0.551	0.607	0.615	**0.623**	0.555	0.607	0.586
Michalewicz	0.798	**0.818**	0.789	0.574	0.749	0.777	0.533	0.741	0.726
Rastrigin	0.602	**0.633**	0.622	0.542	0.606	0.630	0.590	0.571	0.627
Rosenbrock	0.976	**0.980**	0.977	0.970	0.966	0.965	0.946	0.950	0.943
SixHumpCamel	0.985	0.990	**0.991**	0.986	0.983	0.975	0.986	0.954	0.951

Table 7.7: *R^2 metric scores for the examined test functions. Bold values indicate the highest score achieved for each respective test function (row), and asterisks (*) denote the overall highest score within each row. It's important to note that due to rounding, some values may appear equal, but only one is the true highest score.*

$$
\begin{aligned}
2A &\rightarrow B \\
B &\rightarrow C \\
C &\rightarrow D \\
B + D &\rightarrow E
\end{aligned}
$$

Table 7.8: *Syntethic model. Four reactions, five species.*

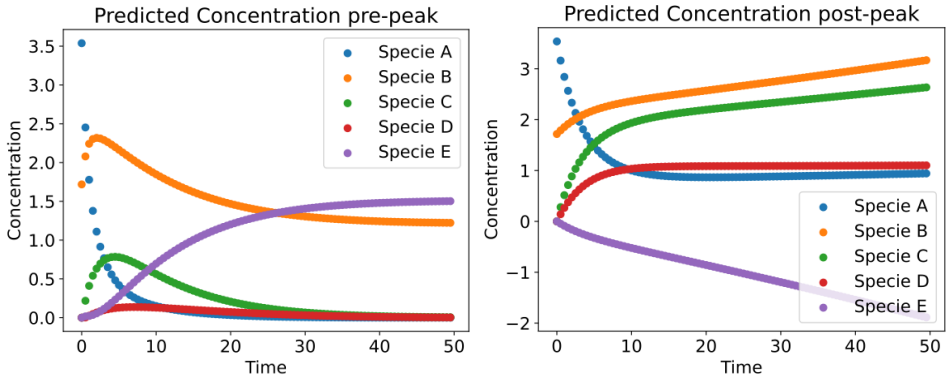
ference between the learned model before (Table 7.9) and after (Table 7.10) consists in the consumption of the species E, even if it is never generated and its initial concentration is 0. Therefore, during the learning procedure, the model is proposing a unphysibale solution in which it is consuming a species that does not exist, violiting the physic laws.

7.3.1 Element Conservation

A possible solution to ensure that the learned predictive model does not violate the physical laws is to impose element conservation in each reaction.

$$
\begin{aligned}
0.1B + 0.9C &\rightarrow 0.6A + 0.1D + 0.4E \\
0.7A + 0.1B + 0.1E &\rightarrow 0.5C + 0.1D \\
2.9A + 0.1D &\rightarrow 1.4B + 0.1C + 0.2E \\
0.9B + 0.9D &\rightarrow 0.4C + 0.5E
\end{aligned}
$$

Table 7.9: *Stochiometric coefficients before the loss function peak.*



(a) **(b)**

Figure 7.15: *CRNN prediction of a given set of experimental data (a) before the peak and (b) after the peak in the training loss function Figure 7.14.*

$$
\begin{aligned}
0.6C + 0.1E &\rightarrow 0.6A + 0.2B + 0.4D \\
1.1A + 0.1E &\rightarrow 0.2B + 0.9C + 0.4D \\
2.6A + 0.5D + 0.3E &\rightarrow 0.9B + 0.2C \\
0.5B + 1.1D &\rightarrow 0.2A + 0.4C
\end{aligned}
$$

Table 7.10: *Stochiometric coefficients after the loss function peak.*

In other words, the number of elements generated and consumed in each reaction must be preserved between the reaction reactants and products. In the literature, there are two possible ways to enforce constraint in a neural network [181]. The first one (Hard constraint) is enforcing the constraint in the NN architecture [229, 230], the second (Soft constraint) is to penalizes the prediction of the NN does not fulfill the constraint [231–233]. The second case is also known as Physically Informed Neural Network (PINN).

However, to the best of my knowledge, nobody has enforced the element conservation in a NN. Therefore, this thesis proposes the following methodology to ensure element conservation as a soft constraint.

A species A can be composed of multiple elements. For instance, water, H_2O, is composed of two elements of hydrogen H and one of oxygen O.

Given a model, it is possible to construct its stochiometric matrix $S_M = |S| \times |R|$ where each cell specifies the stochiometric coefficient of the species $i-$th in the $j-$th reaction.

For instance, given the model in Table 7.8, the S_M is $\begin{bmatrix} -2 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.
The reactants have negative values, whereas they are positive for the products. Similarly, it is possible to define an element composition matrix $E_M = |S| \times |E|$, where E is the set of the existing elements in the model. Examples of elements are hydrogen H, Oxygen O, and Carbon C. For instance if $S = \{H_2O; CO_2; H_2\}$ and $E = \{H, O, C\}$, the corresponding $E_M = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 1 & -1 & 0 \end{bmatrix}$

The element loss $LossE_j$ for each reaction in S_M is computed considering the $j-$th column as follows in Equation (7.13). The element-stochiometric matrix is defined as $ES = S_M \circ (R_j \times 1^{|S|})$, where $1^{|S|}$ is row-vector of ones of dimension $|D|$. α is a correction factor to normalize the magnitude of the learned stochiometric coefficient. Two different α can be defined, one for the reactants and one for the products, such as at least one reactant, and one product has at least a stochiometric coefficient of 1.

$$LossE_j = \frac{\sum_{j=1}^{|E|} |\sum_{i=1}^{|S|} ES_{ij}\alpha|}{|E|} \tag{7.13}$$

Finally, the total element loss of a model is the sum of the single element loss of each reaction (Equation (7.14)).

$$LossE = \sum_{j=1}^{|R|} LossE_j \tag{7.14}$$

Future works will investigate the application of such element loss in CRNN. However, so far, different ideas have been investigated to evaluate and develop a scientific predictive model objectively, but it is also important to investigate some ethical considerations of such procedures and the possible mitigations.

7.4 Data Ethics

Predictive models are increasingly persuasive in every area of daily life, from engineering to the social sciences. Their use is gradually going to automate and replace areas that were typically done or are no longer sustainable manually. Recently, their deployment has been facilitated by the increasing amount of data on which it is possible to build predictive models using a data-driven approach, often using data science techniques.

The first wave of data science aimed to improve predictive models in terms of accuracy and efficiency. However, all the ethical implications and the careless use of these models have been overlooked. After several ethical problems arose during the second wave of data science, the focus shifted from what could be done with data to what and how we should or should not do with it. As a result of this new ethical attention to the use of the data, methodologies were proposed to analyze the entire life cycle of the data. Data Quality (DQ) is one of the main factors that are often under the spotlight. It is fundamental to build an accurate and ethical predictive model since it directly affects the model's outcomes. However, other data-related aspects are also important, such as diversity and provenance.

This section discusses that such aspects are also important and should be regularly treated in the data science ethical-technical debate. Therefore, starting from practical examples, the following first presents data quality, diversity, and provenance problems. Then, it discusses the corresponding trade-offs and mitigations and how they coexist and cooperate to address ethical issues from a technical perspective.

In 2006, British mathematician Clive Humby used for the first time the expression "Data is the new oil." That phrase, which at the time might have seemed almost like a provocation, we know today that phrase is a reality. From that time, there has been a real *"data-oil" rush*, in which we have witnessed the use of terms such as big data, data analytics, and artificial intelligence entering the ubiquity of everyday life. Universities have also adapted to this phenomenon. They started to propose entire degrees on these topics to meet the growing demand for data-related positions. Companies, sometimes also dazzled by the promising results of the employment of a large amount of data and to maintain a certain appeal with the product's market, push hard to launch applications of this kind or create job positions that include these main-stream terms in the name.

Artificial intelligence and, more in general, data science-related topics have brought both positive and negative benefits. Indeed, due to the numerous scandals, it is evident that there is a lack of responsible develop-

ment [234] creating even more accentuated social inequalities [235]. As a countermeasure, the public sector is increasingly forced to pursue ethical policy in its algorithms, constantly under scrutiny and criticism [235].

Nowadays, artificial intelligence capabilities are not in doubt. In some fields and tasks, they even overpower human capabilities, although how to actually evaluate these capacities is still a matter of debate [236]. To achieve such levels, there is a need to manage, process, and analyze large amounts of data. Then a data scientist uses the data to create a predictor. A predictor is a tool that predicts a not known situation based on what they have seen in the past. In these terms, the predictor has learned patterns in the data, hence the term “machine learning". They are difficult-to-interpret tools that are gradually supplanting many human tasks, even of social and ethical significance, for their fast ability to perform tasks.

For these promising prospects, data scientists are often under pressure to deliver a product quickly, such as a predictive model. However, the rapid and, therefore, often uncontrolled development implies that they are not considering aspects such as the ethical implications. These implications are due to how the product has been used or built [237]. In other words, data scientists often have no idea of the power they have by developing these types of products, so they act careless and indifferent to these types of problems and their possible consequences [238]. Other scholars, on the other hand, believe that this technical-social gap is not due to carelessness but rather to the incredible complexity of the problems to be addressed [238]. In her book *Weapons of Math Destruction* Cathy O’Neil points out that data scientists need to recognize how big companies use their skills to achieve business goals without thinking about the consequences [239], even if it is still under discussion [238]. What is not disputed is that given the obvious problems and the active ethical and not ethical discussion, there is a regulatory emptiness, and the existing frameworks are not troubleshooters for practitioners [234, 235, 238]. It is fundamental to cover this gap, including in the discussion both ethical and technical perspectives and include all kinds of persons involved in the design, delivery, and employment of a data-driven product [235, 240].

Nowadays, there is a need to process very large amounts of data with fewer and fewer resources, including human resources. The only solution to keep the rhythm is to employ automatic tools where possible, and in the activities where there is a social or technological challenge (for now) in replacing the human, you flank it with an intelligent component. In the meanwhile, these algorithms are becoming much more complex and challenging to explain and interpret the results. Instead, it was easier in the past

since they were also designed for simpler tasks. What is left is the trust in algorithms, and technology in general, that can perform a human task fast, precisely, and reliably.

The development of predictive models is the result of a long pipeline of tasks, where data is processed, and it is the main protagonist. These tasks vary from data collection and preparation to annotation and visualization. Ethical requirements can be viewed in terms of DQ dimensions [241]. So that only data that meet ethical requirements are to be considered of high quality. Starting with DQ, this chapter describes, from a technical perspective, two other critical ethical aspects that can be seen as DQ dimensions: Data Diversity and Data Provenance. In particular, it follows the design principles for ethical research in the era of big data to achieve qualitative research [242]. These topics are critical since they can be translated into fairness, neutrality, and transparency in data analysis [243].

7.4.1 Data Quality

Nowadays, data is being generated and collected at an unprecedented rate. DQ is a fundamental aspect of data management since it directly affects everyday life since many decision-making applications are built on a massive amount of data, also known as Big Data [9, 196, 244].

On the other hand, DQ has been increasingly investigated in recent years. In fact, DQ, with the study and then the definition of several DQ dimensions, objectively quantifies the quality of a data set, checking if the DQ rules associated with a dimension are fulfilled and how often. The most typical DQ dimensions are completeness, consistency, accuracy, uniqueness, and timeliness. There are multiple examples of poor DQ with ethical implications in the literature.

In 2000, the US Presidential election caused quite a stir due to, among other issues, poor DQ of the voter registration database [245]. These data were affected by duplicated registrations (uniqueness), incorrect addresses (accuracy), and missing information (completeness). All these aspects could be identified by a data profiling activity in which the proper DQ dimensions are assessed. Similar problems are registered in the healthcare system [246], supply chain [247] and finance [248], resulting in an estimated cost of around billions of dollars per year [244, 249, 250].

From these examples, it looks straightforward that better DQ implies ethical applications and vice versa. Although the application of the DQ rule and its measurement is objective, the definition of the rules to be checked and which DQ dimension should be investigated is not. In fact, the “fitness

Name	Age	Driving License ID
Tom	22	A1234
Elizabeth		A2321
Ben	25	
Ashlee	7	

Table 7.11: *A toy example about the arbitrariness of the application of data quality rules.*

for use" concept allows each data scientist to decide, based on the situation, which are the DQ dimensions of interest. At the same time, based on a particular domain, there is no unique way of identifying the rules and verifying their criteria.

Table 7.11 is a toy example of the challenges related to DQ. Let us consider completeness as the DQ dimension under investigation. Thus, it has been (more or less arbitrarily) decided that this DQ dimension is important for this scenario and has to be investigated. In simple terms, completeness measures how many cell data entries are missing in a data set. In other words, if the data set can be represented as a table like in Table 7.11, the completeness for each column counts how many empty cells. In this case, the quantification of completeness and its definition are straightforward. However, in practical terms, how to define and apply the rule is not. In Table 7.11, considering the column *Age*, the completeness is three out of four or 75%. Regarding *Driving License ID*, the completeness should be 50%. If the data is observed more carefully, “Ashlee" is only seven, thus can not have a *Driving License ID*. Therefore, her corresponding *Driving License ID* cell is not empty. Thus it is not incomplete. Table 7.11 reports a simple case to handle, but, in general, it requires a deep knowledge of the dataset's domain, and a user can define more restrictive rules with respect to others and which DQ dimension to investigate. In the end, domain and problem complexity determine subjectivity in the DQ check and, thus, together with the assessment, are fundamental, providing additional details on how the assessment is performed.

7.4.2 Data Diversity

Data diversity is often under the spotlight in the ethical debate but with a different (related) concept: bias. A biased algorithm, predictive model, or dataset causes a high glamor since the gravity of its implications are easily tangible and immediate to comprehend, also by people without a technical background. In simple terms, bias in data occurs when elements of a dataset are overrepresented with respect to others. Therefore, when em-

ployed for data analysis or to build a data-driven predictive model, it results in a prejudiced outcome due to the disproportion of information. In recent years, we have witnessed (unfortunately) several news that have generated buzz due to the ethical consequences of using biased data. Some examples are Florida's recidivism risk system [251], the Amazon-AI recruiting system [252], Pennsylvania child welfare screening tool [253], and Google Translate gender bias [254]. All these applications have in common that they are built on a dataset containing some bias, mainly regarding race or gender. Studies also have shown a "meta" bias in academia. They have highlighted that investigating possible ethical problems in AI-related fields often does not look for misrepresentation regarding minority groups, such as disabled people [255]. Learning from these experiences, it seems almost easy to split into datasets with bias. Enhancing the diversity in a dataset should bring benefits in two ways: first, to mitigate the risk of exclusion of an underrepresented or overrepresented group for the ethical debate, and second, to gain engagement in building more powerful and accurate applications [165].

Also in this case, assessing the heterogeneity (or coverage) of a dataset is a time-consuming task; therefore, it is fundamental to employ a semi-automatic approach that synthetically summarizes whether the ethical requirements are fulfilled. One of the proposed solutions regards the quantification of the database coverage, in which, after identifying the dimension that characterizes a domain, a corresponding multidimensional matrix counts the number of data for each possible combination in the dataset [40]. More precisely, each data entry in the dataset, for each characteristic or property of the domain, assumes a value. Hence, the matrix dimensions correspond to the identified characteristics of the domain. Each dimension expects to be able to assume a given set of possible values, which will therefore correspond to the cardinality of the dimension itself. As a result, the final user has a qualitative and quantitative overview of the diversity of the dataset. The solution also proposes a synthetic index that summarizes the coverage of the dataset as a ratio between the number of accounted combinations with at least k cases over the total number of the existing combinations in a domain where k is given.

A toy example in Figure 7.16 represents the visualization of this concept: the domain is characterized by three dimensions or characteristics: "Age", "Eye Color", and "Height". Each dimension can assume a precise set of possible values. For instance, in the case of "Eye Color", the values are "Blue" and "Green". In this example, the total number of possible combinations is eight. A diverse dataset that fulfills ethical requirements in

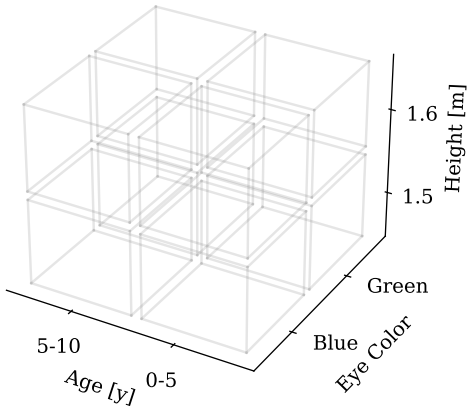


Figure 7.16: *A visualization of the assessment of a data diversity methodology.*

terms of bias should include at least one element in each *box* (or *bucket*), i.e., a combination of property values for each dimension, and, in the general case, not *boxes* that are over or under-represented with respect to others. However, as in the case of data quality, there are some drawbacks. The index, and more in general, depending on how it is used, could result in misleading and unethical results. Considering the example in Figure 7.16 with respect to the case of “Eye Color". The multidimensional matrix accounts only for the “Blue" and “Green" colors, but not, for example, “Brown". Therefore, if the possible values of a domain dimension are not present, or a dimension is not present at all, or the range of a bucket is improperly generous, such as in the case of Age, there is the risk of having a high index score but without having in practice a diverse database. Solutions, like the multidimensional matrix, are fundamental to processing large amounts of data in the Big Data era, but it is also fundamental to properly set the dimensions and their possible values. Data Provenance, in the next section, will be fundamental to overcoming these limitations. Suppose every scientific and non-scientific community can identify the diversity dimension and their possible values uniquely. In that case, data diversity can be measured with this methodology, thus assessing ethical requirements.

7.4.3 Data Provenance

DQ and Data Diversity are two countermeasures that attempt to reconcile the need to process large amounts of data and to check that ethical requirements are met. However, these approaches, for how they are defined to digest “Big" Data, generally summarize the results in performance indexes. Unfortunately, due to the domain complexity in which they are applied, there is no unique way to define the indexes and how they are computed. Therefore, these metrics could be subjective and thus responsible for misleading the fulfillment of ethical requirements. Data transparency can mitigate these limitations. In the literature, data transparency is considered not an ethical principle in itself, but rather it is a pro-ethical prerequisite to enable other ethical practices or principles [256]. In fact, Data transparency, in practice, consists in providing a set of information describing the origin and process to which the data has been subjected. The data record that contains this information is called the Data Provenance or data lineage.

Recently there has been increasing recognition of the importance of Data Provenance in data-sensitive applications such as artificial intelligence [257]. In other terms, the employment of provenance data is fundamental in critical operations where there are several ethical hazards to establishing and ensuring trust [258]. Trustworthiness must be shown at each level of employment of the data and in terms of who uses it. From the raw data, the methodology, the analysts, and the organizations that offer services from the data. Each of them must show to be trustworthy and be able to decide to use trustworthy elements. Data Provenance can be employed to keep track of the adequacy of the process, i.e., the chain of trust [259]. Future users of a service or data can redeem the trust by being allowed to reproduce every single step [204]. In general, Data Provenance is not only helpful in replicating results but also in tracking down errors and accountability [205].

In practical terms, Data provenance is a collection of metadata that contains information about entities, activities, and users involved in producing, transforming, or using data. Together with the provenance metadata, the provenance data model has to be provided since it is fundamental to understanding what is stored in the provenance record. In fact, the provenance data model defines the schema of the provenance metadata in a domain by relating the various entities, activities, and people involved in the data creation or in the following data processing stages. According to the literature, the design of the provenance data model should follow the data sheets directives [206]. They recommend representing what is strictly necessary for the data preparation pipeline design. In fact, a provenance data model can

have different levels of granularity [260], and thus different verbosity. In any case, all the information to replicate the steps from the data source to the final deployment of the data has to be present.

However, as in the case of data quality and data diversity, also in Data Provenance, there is the human element, thus a subjective factor. Data Provenance and hence trustworthiness depends on the definition of the model that is used to keep track of the relevant aspects of the world [259]. In other words, there is no unique definition of what is relevant to be reported on what is not. In the end, we need to trust the analysts or the data providers, and their training and accreditation are clearly important [259]. In any case, regardless of what is actually stored inside Data Provenance, the Data Provenance record enables a stronger assurance of the data since "more eyes mean fewer lies, and fewer mistakes" [259].

The W3C PROV data model is often used to design the provenance data model [207]. Figure 7.17 represents the W3C PROV data model, i.e., the conceptual provenance model that is built around three elements: *Entity*, *Activity*, and *Agent* [207]. An *Entity* is something for which we want to trace the provenance. It could be a single entry in a dataset, like an image or the entire dataset, or a single document. An *Activity* is an operation performed on an entity to produce another entity or another version of it. Typical activities could be the translation of a document into another language or the cropping or tagging of an image. Finally, the *Agent* is something or someone that bears the responsibility for an action or an entity. An example could be the annotator of a set of images or the data analyst that made a particular decision based on some observations. Each element in the W3C PROV data model is connected to another one by a pre-defined set of relationships as shown in Figure 7.17.

7.4.4 Discussion

Nowadays, predictive models, or, more in general, algorithms that process large amounts of information, are fundamental to continuously offer services with limited resources. In particular, an AI component brings two benefits: the first is the speeding up, and the second is transferring the responsibility. Today, it is possible to develop an intelligent agent to make challenging decisions regarding social aspects of daily life. Even if it could sound scary to rely, for instance, on an algorithm for a health diagnosis or a court judgment, people are getting more used to this practice: relying on technology to replace manual human tasks. From decades of technological improvements, we have developed an unconscious trust in the technology

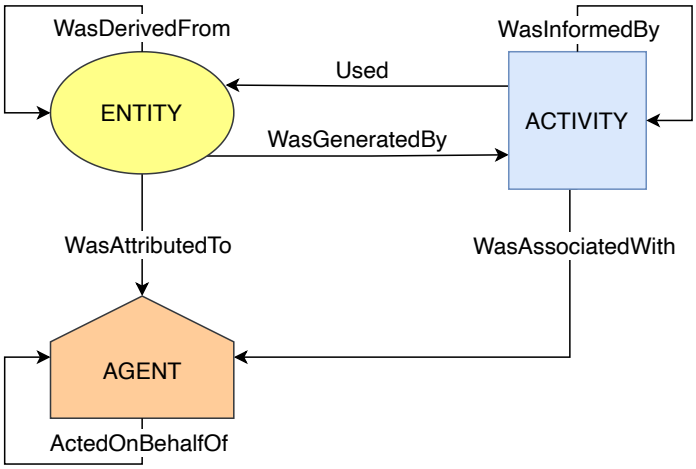


Figure 7.17: *The W3C PROV Data Model.*

being able to deliver the result faster and reliably. However, until now, technology was replacing mainly repetitive and easy tasks, whose outcomes were easy to understand and verify. Today, algorithms are very articulated since they try to mimic human reasoning. However, due to their complexity, it is not possible anymore to easily verify them, but it remains the same trust and expectations in technology.

Due to the increasing demand for these promising tools, product designers and developers are often under pressure to deliver them. As a result, there are no resources to analyze in deep the implemented product from an ethical perspective leading to ethical issues such as discrimination or prejudice. These gaps have great media resonance, both because of the implications and because these products are used in many aspects of daily life. Most of the time, the problem is not the algorithm itself but the “corrupted" data fed into it. Recent technologies, such as neural networks, “just" learn the hidden patterns in the data. Thus, if, for example, the data is biased, the algorithm is. A dataset could represent more or less a particular social activity, meaning that the algorithm is probably biased because the society is. However, when designing and developing an algorithm or a predictive model, there is the opportunity to measure and mitigate ethical issues. It should be easier than changing unethical behaviors in our society. Sometimes, it happens the opposite. The unethical behavior is enlarged when delivering a data-driven product. In both cases, it is fundamental to be resilient to data management-related ethical issues accounting for Data Quality, Data Diversity, and Data Provenance.

This chapter discusses these data management aspects as mitigation for ethical issues from a technical perspective, even if the ethical debate per se is still ongoing, and foresees many different perspectives. It first presents the problems in terms of examples and then corresponding solutions. The mitigations are introduced keeping in mind that we should address the ethical issues in a way that is still feasible to process large amounts of data. On the other hand, providing automatic or semi-automatic tools that do not act unethically is very challenging. It could be opened an ethical debate about whether it is better to use these tools to assess ethical requirements but not totally reliably or not to use them at all. As discussed in this chapter, the best thing we can do is create awareness in who will use a product or data. Data Quality, Diversity, and Provenance are scalable techniques to mitigate technical-ethical issues, but they come with limitations. DQ ensures that the garbage-in garbage-out paradigm of a data-driven model is not reached, ensuring quality data for building these tools. However, which DQ dimension to investigate and how to apply it is subjective. Similarly, Data Diversity measures a dataset's heterogeneity to create awareness of the bias in the data and, thus, in the algorithms. As before, it is not always objective on how to assess it. In the end, Data Provenance has the capability to overcome the previous limitations. In principle, it records, using a given model, all the actions that happened to an entity. Therefore, it can be used to track all the aspects of the design and creation of a predictive model. Unfortunately, also in this case, simply providing the provenance record is not sufficient. The provenance data tracks only what is specified in the provenance data model. In this setting, the origin of the ethical problem in the big data era could have a new different perspective: it is currently missing the engineering part in the assembly of a long and complex pipeline of a product, such as an algorithm, a predictive model, or an analysis report. Due to other constraints, data scientists often just assemble various components of a data pipeline without reasoning about the drawbacks of such linking. On the other hand, the profession of the engineer should be characterized by the ability to understand when and how it is safe to connect various parts of a complex system. In this case, the engineers, after being properly trained, should consider the provenance record and make conscious ethical decisions in their product design.

CHAPTER 8

Discussion and Conclusion

This thesis investigates, as a whole, different aspects of the development process of scientific predictive models. Its goal was to facilitate such research workflow in terms of automation, speed-up, and effectiveness. Therefore, it focuses on applying and developing specific data science and data management methodologies and a Data Ecosystem (DE) to support such a process. This work considers the chemical engineering domain, in particular, chemical kinetics, as a running scenario (Chapter 3). However, the identification of the requirements and challenges and the corresponding proposed solution are generalizable across many scientific domains in which predictive models are developed.

Chapter 4 contributes to formalizing the scientific data, the properties, and what makes adopting a data-sharing platform in a scientific domain challenging. A DE in scientific domains can bring benefits for three reasons: first, due to the scarcity and cost of scientific data, a DE can enhance data sharing within the research community. Second, since the predictive model development process foresees time-consuming steps and error-prone sequences of tasks on scientific data, the DE can transform the tasks into services that can help automate and speed up research workflows. Finally, collecting and organizing the information in a DE can open new frontiers to

discovering hidden insights in analyzing large amounts of data. Although using a DE in this field promises great results, some challenges may arise and impede successful accomplishment. The challenges to adopting a DE in a scientific domain result from the combination of specific domain requirements from the scientific research community and distinctive properties of scientific data. The literature has examples regarding the design and the implementation of a sharing platform in which the data consumer and producer are distinct entities, whereas, in scientific domains, a user is typically both simultaneously. Moreover, there needs to be more attention on how to make such projects long-term initiatives, thus lacking a sustainability plan from an information system perspective, whereas it is already under study in the business community. The proposed user-trust-data framework identifies the challenges into two macro-categories that threaten the long-term use of a DE: cost and engagement. The challenges are then generalized to be applicable in other scientific domains. Low engagement and high-cost limit the number of users, trust in the platform, and the quantity of shared data. These challenges are drawn from the research experience both in the social media and chemical kinetics domains. In the future, a more quantitative assessment of the analysis dimension could bring benefit in assessing the efficacy of the proposed solution to reach an objective function's optimum. Moreover, since the threats and the causes of such are already investigated in the business literature, mapping them and vice versa in the information system domain would be helpful for both scientific communities.

Chapter 5 proposes the solution to these challenges with the design and implementation of a DE. A DE solution offers both the capabilities of a scientific repository and a collection of services to support and improve the predictive model development process with new functionalities. The literature has examples of centralized, federated, or distributed DE. There is a central data management authority in a centralized or distributed DE. However, the centralized one allows for the independent participation of the DE users, thus not requiring coordination between the participants to keep the DE running with all the available data. A centralized data management system has major control over the data and makes it easier to track the use, misuse, and eventually, the right to be forgotten of data with simpler, less expensive, and binding technologies such as the blockchain. A centralized data control system also allows for more reliable findability. In a distributed DE, even if the data management is centralized, the availability of some functionalities or the entire repository requires coordination and reliable participants. For instance, if one participant went offline, not all

the data could be available to the remaining participants. Therefore, assessing some analyses, such as the data diversity of the repository, could lead to incomplete or unreliable results. A similar situation occurs in a federated DE with no central data management authority. For instance, the data quality policy to accept the data in the repository, such as the required metadata to describe an experiment, could differ based on the participant. In a centralized DE, the availability of services and data does not depend on the participants, and the data management policies are the same among all the participants. On the other hand, this approach is easier to scale up. This work, to conciliate the design principles of DE and to address the previous challenges, proposes a hybrid DE configuration: central data management with federated computational resources. This hybrid architecture is a trade-off between the centrality of an organizational DE that has control over the data but simultaneously encourages and promotes data sharing and the scalability of the system. After identifying and analyzing the existing model development procedure, a new process is defined, introducing the identification and separation of roles and new stages, such as data preparation. All the functionalities are implemented in a DE with a micro-services architecture currently used by multiple research groups. Therefore, this work, starting from the business level, has identified the necessary organizational, architectural, and technological levels going to affect the business level from which it had started by suggesting that some of the steps already present be added, included, or modified. Future works concern a deeper analysis of the possible confidentiality policy in a data-sharing platform, the study of a token strategy to incentivize data sharing and avoid free rides, and the integration of new architecture such as data mesh and data lakes as possible solutions for such thesis objectives.

Chapter 6 focuses on the data preparation aspects of such a predictive model development process. The development of a predictive model is a data-driven activity. Thus, the data has a direct and massive influence on the resulting products. Since the predictive model is the result of a long pipeline, understanding the impact of each phase is not easy. This work proposes a data pipeline design methodology using provenance information to enhance the trustworthiness and improve the pipeline itself. Model validation in the context of a predictive model is highly dependent on the quality, quantity, and diversity of the data used for the validation. Therefore, this thesis also proposes a new methodology to assess the diversity of a dataset, ensure certain data quality in the scientific repository, and predict the missing information. In particular, during model validation, the simulated data by the predictive model are compared with the experimental data

available in the DE repository. However, experimental data are affected by uncertainty, and quite often, this information is missing, even if it is crucial to properly assess the model performance. Given these challenges, this thesis demonstrates the application of knowledge graph embedding to predict the missing uncertainty information. Future developments will concern the analysis of such results using different embedding models and the impact of the knowledge graph topology on the embedding quality. Finally, since knowledge graph embedding has promising results for data quality-related activities, we plan to use it for data cleaning and outlier detection tasks.

Chapter 7 finally proposes a standardized, objective, and systematic model evaluation procedure to understand *why*, *where*, and *how much* the prediction of the model deviates from the experimental data, developing ad-hoc algorithms. The thesis also investigates the ethical impact of the different model development process stages and proposes the corresponding mitigations. Finally, it is presented a novel adaptive sampling algorithm that achieves at the same time good generalization and optimization capabilities, unlike other sampling algorithms, leveraging the generalization capabilities of Neural Network (NN) and the geometric properties of the Delaunay triangulation. Applications of such algorithms vary from the Design of Experiments (DOEs) to selecting the most informative set of training data and, thus, reducing the number of information and resources needed to develop a predictive model. Regarding improving a chemical kinetic model, it is investigating the application of a promising technology named Chemical Reaction Neural Network (CRNN) to develop a hydrogen model. In the training process, this Neural Ordinary Differential Equation (NODE) combines the black-box approach of NNs with well-known physical-chemical laws. After investigating the capabilities of such technology, this work proposes to incorporate element conservation in such architecture. Current challenges and future works are related to numerical issues of CRNN on the hydrogen data and the systematic performance assessment of the proposed adaptive sampling algorithm against the other adaptive sampling algorithms. In the future, the investigation of large language models for the generation of scientific predictive models will be a promising area of research.

Glossary

AI	Artificial Intelligence.
API	Application Programming Interface.
BPMN	Business Process Model and Notation.
CDRC	Consumer Data Research Centre.
CM	Curve Matching.
CRNN	Chemical Reaction Neural Network.
CSV	Comma-separated Values.
DE	Data Ecosystem.
DOE	Design of Experiment.
DOI	Digital Object Identifier.
DQ	Data Quality.
EOSC	European Open Science Cloud.
FAIR	Findable, Accessible, Interoperable, and Reusable.
GIGO	Garbage In - Garbage Out.
GPR	Gaussian Process Regressor.
HTTP	Hypertext Transfer Protocol.

JSON	JavaScript Object Notation.
KG	Knowledge Graph.
KGE	Knowledge Graph Embedding.
LHD	Latin Hypercube Design.
LHS	Latin Hypercube Sampling.
LP	Link Prediction.
MADO	Multi Adaptive Delaunay Optimization.
MAE	Mean Absolute Error.
ML	Machine Learning.
MSE	Mean Squared Error.
NMAE	Normalized Mean Absolute Error.
NN	Neural Network.
NODE	Neural Ordinary Differential Equation.
NRMSE	Normalized Root Mean Squared Error.
PCA	Principal Component Analysis.
PINN	Physically Informed Neural Network.
RDF	Resource Description Framework.
RMSE	Root Mean Squared Error.
SciExpeM	Scientific Experiments and Models.
SSE	Sum Squared Error.
XML	Extensible Markup Language.

Bibliography

[1] "Data growth worldwide 2010-2025," https://www.statista.com/statistics/871513/worldwide-data-created/, accessed: 2023-12-18.

[2] K. Shilton, E. Moss, S. A. Gilbert, M. J. Bietz, C. Fiesler, J. Metcalf, J. Vitak, and M. Zimmer, "Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research," *Big Data & Society*, vol. 8, no. 2, p. 20539517211040759, 2021.

[3] A. A. Salamai, "Deep learning framework for predictive modeling of crude oil price for sustainable management in oil markets," *Expert Systems with Applications*, vol. 211, p. 118658, 2023.

[4] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, p. 100605, 2023.

[5] N. Nordin, Z. Zainol, M. H. M. Noor, and L. F. Chan, "An explainable predictive model for suicide attempt risk using an ensemble learning and shapley additive explanations (shap) approach," *Asian journal of psychiatry*, vol. 79, p. 103316, 2023.

[6] C. Acciarini, F. Cappa, P. Boccardelli, and R. Oriani, "How can organizations leverage big data to innovate their business models? a systematic literature review," *Technovation*, vol. 123, p. 102713, 2023.

[7] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[8] E. W. Steyerberg and E. Steyerberg, *Applications of prediction models*. Springer, 2009.

[9] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[10] O. H. Hamid, "Data-centric and model-centric ai: Twin drivers of compact and robust industry 4.0 solutions," *Applied Sciences*, vol. 13, no. 5, p. 2753, 2023.

[11] I. Pan, L. R. Mason, and O. K. Matar, "Data-centric engineering: integrating simulation, machine learning and statistics. challenges and opportunities," *Chemical Engineering Science*, vol. 249, p. 117271, 2022.

[12] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review*, vol. 1, no. 2, pp. 293–314, 2014.

[13] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. De-la Hoz-Franco, and E. De-La-Hoz-Valdiris, "Trends and future perspective challenges in big data," in *Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania*. Springer, 2022, pp. 309–325.

[14] H. Patel, S. Guttula, N. Gupta, S. Hans, R. S. Mittal, and L. N, "A data centric ai framework for automating exploratory data analysis and data quality tasks," *ACM Journal of Data and Information Quality*, 2023.

[15] A. Danandeh Mehr, A. Rikhtehgar Ghiasi, Z. M. Yaseen, A. U. Sorman, and L. Abualigah, "A novel intelligent deep learning predictive model for meteorological drought forecasting," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 10 441–10 455, 2023.

[16] T. Koide, W. Lee Pang, and N. S. Baliga, "The role of predictive modelling in rationally re-engineering biological systems," *Nature Reviews Microbiology*, vol. 7, no. 4, pp. 297–305, 2009.

[17] J. A. Miller, R. Sivaramakrishnan, Y. Tao, C. F. Goldsmith, M. P. Burke, A. W. Jasper, N. Hansen, N. J. Labbe, P. Glarborg, and J. Zádor, "Combustion chemistry in the twenty-first century: Developing theory-informed chemical kinetics models," *Progress in Energy and Combustion Science*, vol. 83, p. 100886, 2021.

[18] A. Dreizler, H. Pitsch, V. Scherer, C. Schulz, and J. Janicka, "The role of combustion science and technology in low and zero impact energy transformation processes," *Applications in Energy and Combustion Science*, vol. 7, p. 100040, 2021.

[19] K. Kohse-Höinghaus, "Combustion, chemistry, and carbon neutrality," *Chemical Reviews*, vol. 123, no. 8, pp. 5139–5219, 2023.

[20] D. A. Beck, J. M. Carothers, V. R. Subramanian, and J. Pfaendtner, "Data science: Accelerating innovation and discovery in chemical engineering," *AIChE Journal*, vol. 62, no. 5, pp. 1402–1416, 2016.

[21] T. Davenport, *Big data at work: dispelling the myths, uncovering the opportunities*. Harvard Business Review Press, 2014.

[22] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. De Laat, "Addressing big data challenges for scientific data infrastructure," in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. IEEE, 2012, pp. 614–617.

[23] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on knowledge and data engineering*, vol. 29, no. 10, pp. 2318–2331, 2017.

[24] L. Chiang, B. Lu, and I. Castillo, "Big data analytics in chemical engineering," *Annual review of chemical and biomolecular engineering*, vol. 8, pp. 63–85, 2017.

[25] S. Stall, L. Yarmey, J. Cutcher-Gershenfeld, B. Hanson, K. Lehnert, B. Nosek, M. Parsons, E. Robinson, and L. Wyborn, "Make scientific data fair," *Nature*, vol. 570, no. 7759, pp. 27–29, 2019.

[26] R. A. Poldrack and K. J. Gorgolewski, "Making big data open: data sharing in neuroimaging," *Nature neuroscience*, vol. 17, no. 11, pp. 1510–1517, 2014.

[27] J. Giles, "Scientific uncertainty: When doubt is a sure thing," *Nature*, vol. 418, no. 6897, pp. 476–479, 2002.

[28] D. A. Farber, "Uncertainty," *Geo. LJ*, vol. 99, p. 901, 2010.

[29] I. V. Pasquetto, B. M. Randles, and C. L. Borgman, "On the reuse of scientific data," *Data Science Journal*, vol. 16, no. 8, pp. 1–9, 2017.

[30] R. A. van Santen *et al.*, *Chemical kinetics and catalysis*. Springer Science & Business Media, 2013.

[31] T. Devriendt, P. Borry, and M. Shabani, "Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers," *PLoS One*, vol. 16, no. 7, p. e0254202, 2021.

[32] I. Susha, "Establishing and implementing data collaborations for public good: A critical factor analysis to scale up the practice," *Information Polity*, vol. 25, no. 1, pp. 3–24, 2020.

[33] C. J. Taylor, A. Pomberger, K. C. Felton, R. Grainger, M. Barecka, T. W. Chamberlain, R. A. Bourne, C. N. Johnson, and A. A. Lapkin, "A brief introduction to chemical reaction optimization," *Chemical Reviews*, vol. 123, no. 6, pp. 3089–3126, 2023.

[34] C. Cavallotti, "Automation of chemical kinetics: status and challenges," *Proceedings of the Combustion Institute*, vol. 39, no. 1, pp. 11–28, 2023.

[35] J. Byabazaire, G. O'Hare, and D. Delaney, "Data quality and trust: Review of challenges and opportunities for data sharing in iot," *Electronics*, vol. 9, no. 12, p. 2083, 2020.

[36] E. Ramalli, G. Scalia, B. Pernici, A. Stagni, A. Cuoci, and T. Faravelli, "Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering," *Frontiers in Big Data*, vol. 4, pp. 1–19, 2021, doi: 10.3389/fdata.2021.663410.

[37] E. Ramalli and B. Pernici, "Challenges of a data ecosystem for scientific data," *Data and Knowledge Engineering*, Accepted, doi: 10.1016/j.datak.2023.102236.

[38] ——, "Sustainability and governance of data ecosystems," in *2023 IEEE International Conference on Web Services (ICWS)*. IEEE, 2023, pp. 740–745.

[39] ——, "From a prototype to a data ecosystem for experimental data and predictive models," in *Proc. of the First International Workshop on Data Ecosystems (DEco'22)*. CEUR-WS, 2022, pp. 18–26.

[40] ——, "Know your experiments: interpreting categories of experimental data and their coverage," in *SeaData at VLDB 2021*. CEUR Workshop Proceedings, 2021, pp. 27–33.

[41] ——, "Knowledge graph embedding for experimental uncertainty estimation," *Information Discovery and Delivery*, ahead-of-print, 2023, doi: 10.1108/IDD-06-2022-0060.

[42] C. Bono, M. O. Mülâyim, C. Cappiello, M. J. Carman, J. Cerquides, J. L. Fernandez-Marquez, M. R. Mondardini, E. Ramalli, and B. Pernici, "A citizen science approach for analyzing social media with crowdsourcing," *IEEE Access*, vol. 11, pp. 15 329–15 347, 2023, doi: 10.1109/ACCESS.2023.3243791.

[43] C. A. Bono, C. Cappiello, B. Pernici, E. Ramalli, and M. Vitali, "Pipeline design for data preparation for social media analysis," *ACM Journal of Data and Information Quality*, 2022, doi: 10.1145/3597305.

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

[45] E. Ramalli, T. Dinelli, A. Nobili, A. Stagni, B. Pernici, and T. Faravelli, "Automatic validation and analysis of predictive models by means of big data and data science," *Chemical Engineering Journal*, vol. 454, p. 140149, 2023, doi: 10.1016/j.cej.2022.140149.

[46] E. Ramalli, *Next Generation Technology - Designing the Common Good*. Springer, accepted, ch. Data Quality, Data Diversity and Data Provenance: An Ethical Perspective.

[47] M. Dilettis and E. Ramalli, "Multiple adaptive delaunay optimization," ., In preparation.

[48] E. Ramalli, B. Pernici, T. Faravelli, and D. Sili, "Chemical reaction neural network with element conservation for hydrogen model." ., In preparation.

[49] D. Delen and H. Demirkan, "Data, information and analytics as services," pp. 359–363, 2013.

[50] C. Allan *et al.*, "Omero: flexible, model-driven data management for experimental biology," *Nature Methods*, vol. 9, no. 3, pp. 245–253, 2012.

[51] K. Rafes and C. Germain, "A platform for scientific data sharing," in *BDA2015-Bases de Données Avancées*, 2015.

[52] G. Scalia, M. Pelucchi, A. Stagni, A. Cuoci, T. Faravelli, and B. Pernici, "Towards a scientific data framework to support scientific model development," *Data Sci.*, 2019.

[53] S. Scheider, F. Lauf, S. Geller, F. Möller, and B. Otto, "Exploring design elements of personal data markets," *Electronic Markets*, vol. 33, no. 1, pp. 1–16, 2023.

[54] I. Jussen, J. Schweihoff, V. Dahms, F. Möller, and B. Otto, "Data sharing fundamentals: Definition and characteristics," in *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2023, pp. 3685–3694.

[55] F. Berman, R. Rutenbar, B. Hailpern, H. Christensen, S. Davidson, D. Estrin, M. Franklin, M. Martonosi, P. Raghavan, and V. Stodden, "Realizing the potential of data science," *Communications of the ACM*, vol. 61, no. 4, April 2018.

[56] F. Casati, S. Ceri, B. Pernici, and G. Pozzi, "Conceptual modeling of workflows," in *OOER'95: Object-Oriented and Entity-Relationship Modeling: 14th International Conference Gold Coast, Australia, December 13–15, 1995 Proceedings 14*. Springer, 1995, pp. 341–354.

[57] B. Otto, "A federated infrastructure for European data spaces," *Communications of the ACM*, vol. 65, no. 4, pp. 44–45, 2022.

[58] N. Vij, "Introducing the consumer data research centre (cdrc)," *Journal of Direct, Data and Digital Marketing Practice*, vol. 17, pp. 232–235, 2016.

[59] M. I. S. Oliveira and B. F. Lóscio, "What is a data ecosystem?" in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 2018, pp. 1–9.

[60] G. Alonso, F. Casati, H. Kuno, V. Machiraju, G. Alonso, F. Casati, H. Kuno, and V. Machiraju, *Web services*. Springer, 2004.

[61] E. Curry and A. P. Sheth, "Next-generation smart environments: From system of systems to data ecosystems," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 69–76, 2018. [Online]. Available: https://doi.org/10.1109/MIS.2018.033001418

[62] V. Stodden, "The data science life cycle: a disciplined approach to advancing data science as a science," *Commun. ACM*, vol. 63, no. 7, pp. 58–66, 2020. [Online]. Available: https://doi.org/10.1145/3360646

[63] Y. Cui, S. Kara, and K. C. Chan, "Manufacturing big data ecosystem: A systematic literature review," *Robotics and computer-integrated Manufacturing*, vol. 62, p. 101861, 2020.

[64] M. L. Brodie, "Data integration at scale: From relational data integration to information ecosystems," in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE, 2010, pp. 2–3.

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

[66] W. Fan and F. Geerts, "Foundations of data quality management," *Synthesis Lectures on Data Management*, vol. 4, no. 5, pp. 1–217, 2012.

[67] B. Otto, S. Lohmann, S. Auer, G. Brost, J. Cirullies, A. Eitel, T. Ernst, C. Haas, M. Huber, C. Jung *et al.*, *Reference architecture model for the industrial data space*. Fraunhofer-Gesellschaft, 2017. [Online]. Available: https://publica.fraunhofer.de/handle/publica/298818

[68] B. Otto and M. Jarke, "Designing a multi-sided data platform: findings from the international data spaces case," *Electronic Markets*, vol. 29, no. 4, pp. 561–580, 2019.

[69] Y. Demchenko, C. De Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," in *2014 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2014, pp. 104–112.

[70] B. Otto, M. t. Hompel, and S. Wrobel, "International data spaces: Reference architecture for the digitization of industries," *Digital transformation*, pp. 109–128, 2019.

[71] C. Cappiello, A. Gal, M. Jarke, and J. Rehof, "Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391)," *Dagstuhl Reports*, vol. 9, no. 9, pp. 66–134, 2020. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2020/11845

[72] T. Berlage, C. Claussen, S. Geisler, C. A. Velasco, and S. Decker, "Medical data spaces in healthcare data ecosystems," in *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer International Publishing Cham, 2022, pp. 291–311.

[73] V. Janev, M.-E. Vidal, D. Pujić, D. Popadić, E. Iglesias, A. Sakor, and A. Čampa, "Responsible knowledge management in energy data ecosystems," *Energies*, vol. 15, no. 11, p. 3973, 2022.

[74] J. Gelhaar and B. Otto, "Challenges in the emergence of Data Ecosystems," in *Pacific Asia Conference on Information Systems*, 2020, p. 175.

[75] D. Lis and B. Otto, "Data governance in data ecosystems–insights from organizations," in *Proc. AMCIS*, 2020, p. 20.

[76] L. Özcan, C. Koldewey, E. Duparc, H. van der Valk, B. Otto, and R. Dumitrescu, "Why do digital platforms succeed or fail? – A literature review on success and failure factors," in *Proc. AMCIS*, 2022, p. 15.

[77] B. Otto, "The evolution of data spaces," in *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer International Publishing Cham, 2022, pp. 3–15.

[78] E. Curry and A. Sheth, "Next-generation smart environments: From system of systems to data ecosystems," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 69–76, 2018.

[79] M. Frenklach, "Transforming data into knowledge—process informatics for combustion chemistry," *Proceedings of the Combustion Institute*, vol. 31, no. 1, pp. 125–140, 2007.

[80] G. L. Goteng, N. Nettyam, and S. M. Sarathy, "CloudFlame: Cyberinfrastructure for combustion research," in *2013 International Conference on Information Science and Cloud Computing Companion*. IEEE, 2013, pp. 294–299.

[81] F. Bartolomucci and G. Bresolin, "Fostering data collaboratives' systematisation through models' definition and research priorities setting," in *DG.O 2022: The 23rd Annual International Conference on Digital Government Research*, ser. dg.o 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 35–40. [Online]. Available: https://doi.org/10.1145/3543434.3543442

[82] I. Susha, T. van den Broek, A.-F. van Veenstra, and J. Linåker, "An ecosystem perspective on developing data collaboratives for addressing societal issues: The role of conveners," *Government Information Quarterly*, vol. 40, no. 1, p. 101763, 2023.

[83] J. Schweihoff, I. Jussen, V. Dahms, F. Möller, and B. Otto, "How to share data online (fast)–a taxonomy of data sharing business models," in *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2023.

[84] E. Ruijer, "Designing and implementing data collaboratives: A governance perspective," *Government Information Quarterly*, vol. 38, no. 4, p. 101612, 2021.

[85] B. Otto and K. Weber, "Data governance," *Daten-und Informationsqualität: Auf dem Weg zur Information Excellence*, pp. 277–295, 2011.

[86] I. Susha, Å. Grönlund, and M. Janssen, "Organizational measures to stimulate user engagement with open data," *Transforming Government: People, Process and Policy*, vol. 9, no. 2, pp. 181–206, 2015.

[87] I. Susha, B. Rukanova, A. Zuiderwijk, J. R. Gil-Garcia, and M. G. Hernandez, "Achieving voluntary data sharing in cross sector partnerships: Three partnership models," *Information and Organization*, vol. 33, no. 1, p. 100448, 2023.

[88] S. P. Karimireddy, W. Guo, and M. I. Jordan, "Mechanisms that incentivize data sharing in federated learning," *arXiv preprint arXiv:2207.04557*, 2022.

[89] S. S. Tay, X. Xu, C. S. Foo, and B. K. H. Low, "Incentivizing collaboration in machine learning via synthetic data rewards," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 9448–9456.

[90] M. Jarke, B. Otto, and S. Ram, "Data sovereignty and data space ecosystems," *Bus. Inf. Syst. Eng.*, vol. 61, no. 5, pp. 549–550, 2019. [Online]. Available: https://doi.org/10.1007/s12599-019-00614-2

[91] T. B. Nguyen, K. H. Bates, R. S. Buenconsejo, S. M. Charan, E. E. Cavanna, D. R. Cocker III, D. A. Day, M. P. DeVault, N. M. Donahue, Z. Finewax *et al.*, "Overview of icarus- a curated, open access, online repository for atmospheric simulation chamber data," *ACS Earth and Space Chemistry*, 2023.

[92] T. Nacházel, F. Babič, M. Baiguera, P. Čech, M. Husáková, P. Mikuleckỳ, K. Mls, D. Ponce, D. Salmanidou, K. Štekerová *et al.*, "Tsunami-related data: A review of available repositories used in scientific literature," *Water*, vol. 13, no. 16, p. 2177, 2021.

[93] B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, and I. Foster, "A data ecosystem to support machine learning in materials science," *MRS Communications*, vol. 9, no. 4, pp. 1125–1133, 2019.

[94] C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett, "Changes in data sharing and data reuse practices and perceptions among scientists worldwide," *PLOS ONE*, vol. 10, no. 8, pp. 1–24, 08 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0134826

[95] H. Piwowar and T. Vision, "Data reuse and the open data citation advantage," *PeerJ*, vol. 1, p. e175, 10 2013.

[96] G. George, E. Osinga, D. Lavie, and B. Scott, "Big data and data science methods for management research," *The Academy of Management Journal*, vol. 59, pp. 1493–1507, 10 2016.

[97] M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, and B. Mons, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 03 2016.

[98] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the meaningfulness of "big data quality"," *Data Science and Engineering*, vol. 1, pp. 6–20, 2016.

[99] J. Galvão, A. Leon, C. Costa, M. Y. Santos, and Ó. P. López, "Towards designing conceptual data models for big data warehouses: The genomics case," in *European, Mediterranean, and Middle Eastern Conference on Information Systems*. Springer, 2020, pp. 3–19.

[100] A. León and O. Pastor, "Enhancing precision medicine: A big data-driven approach for the management of genomic data," *Big Data Research*, vol. 26, p. 100253, 2021.

[101] A. L. Palacio and Ó. P. López, "From big data to smart data: A genomic information systems perspective," in *2018 12th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2018, pp. 1–11.

[102] A. Hegde, W. Li, J. Oreluk, A. Packard, and M. Frenklach, "Consistency analysis for massively inconsistent datasets in bound-to-bound data collaboration," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 6, no. 2, pp. 429–456, 2018.

[103] R. Feeley, P. Seiler, A. Packard, and M. Frenklach, "Consistency of a reaction dataset," *The Journal of Physical Chemistry A*, vol. 108, no. 44, pp. 9573–9583, 2004.

[104] X. You, A. Packard, and M. Frenklach, "Process informatics tools for predictive modeling: Hydrogen combustion," *International Journal of Chemical Kinetics*, vol. 44, no. 2, pp. 101–116, 2012.

[105] M. Frenklach, A. Packard, and P. Seiler, "Prediction uncertainty from models and data," in *Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301)*, vol. 5. IEEE, 2002, pp. 4135–4140.

[106] Z. Reyno-Chiasson, N. Nettyam, G. Goteng, M. Speight, B. Lee, S. Baskaran, J. Oreluk, A. Farooq, H. Im, M. Frenklach *et al.*, "Cloudflame and prime: accelerating combustion research in the cloud. 9th in-ternational conference on chemical kinetics," *Ghent, Belgium*, 2015.

[107] T. Varga, T. Turányi, E. Czinki, T. Furtenbacher, and A. Császár, "ReSpecTh: a joint reaction kinetics, spectroscopy, and thermochemistry information system," in *Proceedings of the 7th European Combustion Meeting*, vol. 30, 2015, pp. 1–5.

[108] C. Olm, I. G. Zsély, T. Varga, H. J. Curran, and T. Turányi, "Comparison of the performance of several recent syngas combustion mechanisms," *Combustion and Flame*, vol. 162, no. 5, pp. 1793–1812, 2015.

[109] H. Gossler, L. Maier, S. Angeli, S. Tischer, and O. Deutschmann, "Carmen: an improved computer-aided method for developing catalytic reaction mechanisms," *Catalysts*, vol. 9, no. 3, p. 227, 2019.

[110] V. R. Lambert and R. H. West, "Identification, correction, and comparison of detailed kinetic models," in *9th US Natl Combust Meeting*, 2015.

[111] N. J. Killingsworth, M. J. McNenly, R. A. Whitesides, and S. W. Wagnon, "Cloud based tool for analysis of chemical kinetic mechanisms," *Combustion and Flame*, vol. 221, pp. 170–179, 2020.

[112] B. W. Weber and K. E. Niemeyer, "ChemKED: A human-and machine-readable data standard for chemical kinetics experiments," *International Journal of Chemical Kinetics*, vol. 50, no. 3, pp. 135–148, March 2018.

[113] A. Mirzayeva, N. Slavinskaya, U. Riedel, M. Frenklach, A. Packard, W. Li, J. Oreluk, and A. Hegde, "Investigation of dataset construction parameters and their impact on reaction model optimization using PrIMe," in *2018 AIAA Aerospace Sciences Meeting*, 2018, p. 0143.

[114] P. Zhang, I. G. Zsély, M. Papp, T. Nagy, and T. Turányi, "Comparison of methane combustion mechanisms using laminar burning velocity measurements," *Combustion and Flame*, vol. 238, p. 111867, 2022.

[115] C. Olm, I. G. Zsély, R. Pálvölgyi, T. Varga, T. Nagy, H. J. Curran, and T. Turányi, "Comparison of the performance of several recent hydrogen combustion mechanisms," *Combustion and Flame*, vol. 161, no. 9, pp. 2219–2234, 2014.

[116] D. Q. Gbadago, J. Moon, M. Kim, and S. Hwang, "A unified framework for the mathematical modelling, predictive analysis, and optimization of reaction systems using computational fluid dynamics, deep neural network and genetic algorithm: A case of butadiene synthesis," *Chemical Engineering Journal*, vol. 409, p. 128163, 2021.

[117] J.-P. Simonin, "On the comparison of pseudo-first order and pseudo-second order rate laws in the modeling of adsorption kinetics," *Chemical Engineering Journal*, vol. 300, pp. 254–263, 2016.

[118] J. Fermoso, M. V. Gil, C. Pevida, J. Pis, and F. Rubiera, "Kinetic models comparison for non-isothermal steam gasification of coal–biomass blend chars," *Chemical Engineering Journal*, vol. 161, no. 1-2, pp. 276–284, 2010.

[119] M. Kelly, S. Dooley, and G. Bourque, "Toward machine learned highly reduced kinetic models for methane/air combustion," in *Turbo Expo: Power for Land, Sea, and Air*, vol. 84942. American Society of Mechanical Engineers, 2021.

[120] M. Pelucchi, A. Stagni, and T. Faravelli, "Addressing the complexity of combustion kinetics: Data management and automatic model validation," in *Computer Aided Chemical Engineering*. Elsevier, 2019, vol. 45, pp. 763–798.

[121] M. S. Bernardi, M. Pelucchi, A. Stagni, L. M. Sangalli, A. Cuoci, A. Frassoldati, P. Secchi, and T. Faravelli, "Curve matching, a generalized framework for models/experiments comparison: An application to n-heptane combustion kinetic mechanisms," *Combustion and Flame*, vol. 168, pp. 186–203, 2016.

[122] N. R. Council *et al.*, *On the shoulders of giants: New approaches to numeracy*. National Academies Press, 1990.

[123] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.

[124] S. A. Bell, "A beginner's guide to uncertainty of measurement," Centre for Basic, Thermal and Length Metrology, National Physical Laboratory, 2001.

[125] C. C. Aggarwal and S. Y. Philip, "A survey of uncertain data algorithms and applications," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 5, pp. 609–623, 2008.

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

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

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

[129] Y. Xu, X. Fang, X. Li, J. Yang, J. You, H. Liu, and S. Teng, "Data uncertainty in face recognition," *IEEE Transactions on Cybernetics*, vol. 44, no. 10, pp. 1950–1961, 2014.

[130] Y. Kim, J. Huang, S. Emery *et al.*, "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, 2016.

[131] W. Lidwell, K. Holden, and J. Butler, *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, 2010.

[132] W. Dai, S. Cremaschi, H. J. Subramani, and H. Gao, "Estimation of data uncertainty in the absence of replicate experiments," *Chemical Engineering Research and Design*, vol. 147, pp. 187–199, 2019.

[133] R. G. Hills, "Model validation: model parameter and measurement uncertainty," *Journal of Heat Transfer*, vol. 128, no. 4, pp. 339–351, 10 2006.

[134] C. A. Peters, "Statistics for analysis of experimental data," *Environmental engineering processes laboratory manual*, pp. 1–25, 2001.

[135] R. J. Moffat, "Using uncertainty analysis in the planning of an experiment," *Transactions of ASME Journal of Fluids Engineering*, vol. 107, no. 2, 1985.

[136] S.-H. Hsu, S. D. Stamatis, J. M. Caruthers, W. N. Delgass, V. Venkatasubramanian, G. E. Blau, M. Lasinski, and S. Orcun, "Bayesian framework for building kinetic models of catalytic systems," *Industrial & engineering chemistry research*, vol. 48, no. 10, pp. 4768–4790, 2009.

[137] B. M. Wilson and B. L. Smith, "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. 035301, 2013.

[138] C. C. G. Rodríguez and S. Servigne, "Managing sensor data uncertainty: a data quality approach," *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, vol. 4, no. 1, pp. 35–54, 2013.

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

[140] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2014.

[141] R. Coulon, V. Camobreco, H. Teulon, and J. Besnainou, "Data quality and uncertainty in LCI," *The International Journal of Life Cycle Assessment*, vol. 2, no. 3, pp. 178–182, 1997.

[142] M. Chen and B. Plale, "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, 2012.

[143] N. Schuurman and A. Leszczynski, "Ontology-based metadata," *Transactions in GIS*, vol. 10, no. 5, pp. 709–726, 2006.

[144] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs." *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 48, no. 1-4, p. 2, 2016.

[145] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, and M. Kraft, "Ontokin: An ontology for chemical kinetic reaction mechanisms," *Journal of Chemical Information and Modeling*, vol. 60, no. 1, pp. 108–120, 2019.

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

[147] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[148] Y. Dai, S. Wang, N. N. Xiong, and W. Guo, "A survey on knowledge graph embedding: Approaches, applications and benchmarks," *Electronics*, vol. 9, no. 5, p. 750, 2020.

[149] J. N. Fuhg, A. Fau, and U. Nackenhorst, "State-of-the-art and comparative review of adaptive sampling methods for kriging," *Archives of Computational Methods in Engineering*, vol. 28, pp. 2689–2747, 2021.

[150] J. Eason and S. Cremaschi, "Adaptive sequential sampling for surrogate model generation with artificial neural networks," *Computers & Chemical Engineering*, vol. 68, pp. 220–232, 2014.

[151] H. Liu, J. Cai, and Y.-S. Ong, "An adaptive sampling approach for kriging metamodeling by maximizing expected prediction error," *Computers & Chemical Engineering*, vol. 106, pp. 171–182, 2017.

[152] K. Crombecq, E. Laermans, and T. Dhaene, "Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling," *European Journal of Operational Research*, vol. 214, no. 3, pp. 683–696, 2011.

[153] P. Jiang, L. Shu, Q. Zhou, H. Zhou, X. Shao, and J. Xu, "A novel sequential exploration-exploitation sampling strategy for global metamodeling," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 532–537, 2015.

[154] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, pp. 455–492, 1998.

[155] V. Aute, K. Saleh, O. Abdelaziz, S. Azarm, and R. Radermacher, "Cross-validation based single response adaptive design of experiments for kriging metamodeling of deterministic computer simulations," *Structural and Multidisciplinary Optimization*, vol. 48, pp. 581–605, 2013.

[156] C. Q. Lam, "Sequential adaptive designs in computer experiments for response surface model fit," Ph.D. dissertation, The Ohio State University, 2008.

[157] F. Farazi, M. Salamanca, S. Mosbach, J. Akroyd, A. Eibeck, L. K. Aditya, A. Chadzynski, K. Pan, X. Zhou, S. Zhang *et al.*, "Knowledge graph approach to combustion chemistry and interoperability," *ACS Omega*, vol. 5, no. 29, pp. 18 342–18 348, 2020.

[158] N. Liu, J. Wang, S. Sun, C. Li, and W. Tian, "Optimized principal component analysis and multi-state bayesian network integrated method for chemical process monitoring and variable state prediction," *Chemical Engineering Journal*, vol. 430, p. 132617, 2022.

[159] S. Mittal, S. Pathak, H. Dhawan, and S. Upadhyayula, "A machine learning approach to improve ignition properties of high-ash indian coals by solvent extraction and coal blending," *Chemical Engineering Journal*, vol. 413, p. 127385, 2021.

[160] P. P. Plehiers, I. Lengyel, D. H. West, G. B. Marin, C. V. Stevens, and K. M. Van Geem, "Fast estimation of standard enthalpy of formation with chemical accuracy by artificial neural network correction of low-level-of-theory ab initio calculations," *Chemical Engineering Journal*, vol. 426, p. 131304, 2021.

[161] Y. Ouyang, L. A. Vandewalle, L. Chen, P. P. Plehiers, M. R. Dobbelaere, G. J. Heynderickx, G. B. Marin, and K. M. Van Geem, "Speeding up turbulent reactive flow simulation via a deep artificial neural network: A methodology study," *Chemical Engineering Journal*, vol. 429, p. 132442, 2022.

[162] F. H. Vermeire and W. H. Green, "Transfer learning for solvation free energies: From quantum chemistry to experiments," *Chemical Engineering Journal*, vol. 418, p. 129307, 2021.

[163] X. Chen, L. G. Wang, F. Meng, and Z.-H. Luo, "Physics-informed deep learning for modelling particle aggregation and breakage processes," *Chemical Engineering Journal*, vol. 426, p. 131220, 2021.

[164] A. Shokry, S. Medina-González, P. Baraldi, E. Zio, E. Moulines, and A. Espuña, "A machine learning-based methodology for multi-parametric solution of chemical processes operation optimization under uncertainty," *Chemical Engineering Journal*, vol. 425, p. 131632, 2021.

[165] M. Drosou, H. Jagadish, E. Pitoura, and J. Stoyanovich, "Diversity in big data: A review," *Big data*, vol. 5, no. 2, pp. 73–84, 2017.

[166] K. Linka, S. R. S. Pierre, and E. Kuhl, "Automated model discovery for human brain using constitutive artificial neural networks," *Acta Biomaterialia*, vol. 160, pp. 134–151, 2023.

[167] K. Linka and E. Kuhl, "A new family of constitutive artificial neural networks towards automated model discovery," *Computer Methods in Applied Mechanics and Engineering*, vol. 403, p. 115731, 2023.

[168] Q. Li, H. Chen, B. C. Koenig, and S. Deng, "Bayesian chemical reaction neural network for autonomous kinetic uncertainty quantification," *Physical Chemistry Chemical Physics*, vol. 25, no. 5, pp. 3707–3717, 2023.

[169] X. Su, W. Ji, J. An, Z. Ren, S. Deng, and C. K. Law, "Kinetics parameter optimization of hydrocarbon fuels via neural ordinary differential equations," *Combustion and Flame*, vol. 251, p. 112732, 2023.

[170] B. C. Koenig, P. Zhao, and S. Deng, "Accommodating physical reaction schemes in dsc cathode thermal stability analysis using chemical reaction neural networks," *Journal of Power Sources*, vol. 581, p. 233443, 2023.

[171] W. Ji, F. Richter, M. J. Gollner, and S. Deng, "Autonomous kinetic modeling of biomass pyrolysis using chemical reaction neural networks," *Combustion and Flame*, vol. 240, p. 111992, 2022.

[172] C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett, "Changes in data sharing and data reuse practices and perceptions among scientists worldwide," *PLOS ONE*, vol. 10, no. 8, p. e0134826, 2015.

[173] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations," *Science Advances*, vol. 3, no. 4, p. e1602614, 2017.

[174] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[175] H. Wang and D. A. Sheen, "Combustion kinetic model uncertainty quantification, propagation and minimization," *Progress in Energy and Combustion Science*, vol. 47, pp. 1–31, 2015.

[176] J. N. Kutz, *Data-driven modeling & scientific computation: methods for complex systems & big data*. Oxford University Press, 2013.

[177] C. Cappiello, A. Gal, M. Jarke, and J. Rehof, "Data ecosystems: Sovereign data exchange among organizations (dagstuhl seminar 19391)," in *Dagstuhl Reports*, vol. 9:9. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2020.

[178] T. Faravelli, E. Ranzi, A. Frassoldati, A. Cuoci, M. Mehl, M. Pelucchi, A. Stagni, P. Debiagi, L. P. Maffei, A. Bertolino *et al.*, "The CRECK Modeling Group," http://creckmodeling.chem.polimi.it/.

[179] C. Bono, C. Cappielo, M. Dilettis, M. Falconi, B. Pernici, P. Plebani, E. Ramalli, M. Salnitri, C. Sancricca, J. Wang, and M. Vitali, "Data management in information systems: Experience and challenges from preparing and sharing large datasets," in *Proc. of the 2nd Italian Conference on Big Data and Data Science*, 2023.

[180] D. Diran, T. Hoppe, J. Ubacht, A. Slob, and K. Blok, "A data ecosystem for data-driven thermal energy transition: Reflection on current practice and suggestions for re-design," *Energies*, vol. 13, no. 2, p. 444, 2020.

[181] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics–informed neural networks: Where we are and what's next," *Journal of Scientific Computing*, vol. 92, no. 3, p. 88, 2022.

[182] A. Cuoci, A. Frassoldati, T. Faravelli, and E. Ranzi, "Opensmoke++: An object-oriented framework for the numerical modeling of reactive systems with detailed kinetic mechanisms," *Computer Physics Communications*, vol. 192, pp. 237–264, 2015.

[183] J. Antony, *Design of experiments for engineers and scientists*. Elsevier, 2023.

[184] P. Agarwal, M. Tamer, and H. Budman, "Explainability: Relevance based dynamic deep learning algorithm for fault detection and diagnosis in chemical processes," *Computers & Chemical Engineering*, vol. 154, p. 107467, 2021.

[185] M. C. Swain and J. M. Cole, "Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature," *Journal of chemical information and modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[186] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *Ieee Access*, vol. 5, pp. 7776–7797, 2017.

[187] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[188] H. Koers, D. Bangert, E. Hermans, R. van Horik, M. de Jong, and M. Mokrane, "Recommendations for services in a fair data ecosystem," *Patterns*, vol. 1, no. 5, p. 100058, 2020.

[189] G. Scalia, M. Pelucchi, A. Stagni, A. Cuoci, T. Faravelli, and B. Pernici, "Towards a scientific data framework to support scientific model development," *Data Science*, vol. 2, no. 1-2, pp. 245–273, 2019.

[190] W. L. Chang, D. Boyd, and NBD-PWG NIST Big Data Public Working Group, "Nist Big Data Interoperability Framework: Volume 6, Big Data Reference Architecture [Version 2]," 2019.

[191] C. Rodríguez, M. Baez, F. Daniel, F. Casati, J. C. Trabucco, L. Canali, and G. Percannella, "Rest apis: A large-scale analysis of compliance with principles and best practices," in *Web Engineering: 16th International Conference, ICWE 2016, Lugano, Switzerland, June 6-9, 2016. Proceedings 16*. Springer, 2016, pp. 21–39.

[192] N. Alshuqayran, N. Ali, and R. Evans, "A systematic mapping study in microservice architecture," in *2016 IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA)*, 2016, pp. 44–51.

[193] N. Kratzke and P.-C. Quint, "Understanding cloud-native applications after 10 years of cloud computing - a systematic mapping study," *Journal of Systems and Software*, vol. 126, pp. 1 – 16, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121217300018

[194] L. Bradji and M. Boufaida, "A rule management system for knowledge based data cleaning," 2011.

[195] T. Varga, "Optima++ v1.2: A general C++ framework for performing combustion simulations and mechanism optimization," 2020.

[196] S. Garcia, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[197] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini, "Managing data quality in cooperative information systems," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2002, pp. 486–502.

[198] G. Pang, D. Davidson, and R. Hanson, "Experimental study and modeling of shock tube ignition delay times for hydrogen–oxygen–argon mixtures at low temperatures," *Proceedings of the Combustion Institute*, vol. 32, no. 1, pp. 181 – 188, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1540748908000230

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

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

[201] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo, "Knowledge graph embedding for link prediction: A comparative analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 2, pp. 1–49, 2021.

[202] M. Wang, L. Qiu, and X. Wang, "A survey on knowledge graph embeddings for link prediction," *Symmetry*, vol. 13, no. 3, p. 485, 2021.

[203] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "RotatE: Knowledge graph embedding by relational rotation in complex space," in *Proc. ICLR 2019*, 2019.

[204] K. Cranmer, L. Heinrich, R. Jones, D. M. South *et al.*, "Analysis preservation in ATLAS," in *Journal of Physics: Conference Series*, vol. 664. IOP Publishing, 2015, pp. 1–5.

[205] M. Herschel, R. Diestelkämper, and H. B. Lahmar, "A survey on provenance: What for? What form? What from?" *The VLDB Journal*, vol. 26, no. 6, pp. 881–906, 2017.

[206] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[207] K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker *et al.*, "PROV-DM: The PROV data model," *W3C Recomm.*, vol. 14, pp. 15–16, 2013.

[208] A. Gorelik, *The Enteprise Big Data Lake*. O' Reilly, 2019.

[209] A. Asudeh, Z. Jin, and H. Jagadish, "Assessing and remedying coverage for a given dataset," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019.

[210] S. McKinley and M. Levine, "Cubic spline interpolation," *College of the Redwoods*, vol. 45, no. 1, pp. 1049–1060, 1998.

[211] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive data sets*. Cambridge University Press, 2020.

[212] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[213] R. F. Tate, "Correlation between a discrete and a continuous variable. Point-biserial correlation," *The Annals of Mathematical Statistics*, vol. 25, no. 3, pp. 603–607, 1954.

[214] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

[215] D. Dueck, *Affinity propagation: clustering data by passing messages*. Citeseer, 2009.

[216] P. Seiler, M. Frenklach, A. Packard, and R. Feeley, "Numerical approaches for collaborative data processing," *Optimization and Engineering*, vol. 7, no. 4, pp. 459–478, 2006.

[217] D. E. Edwards, D. Y. Zubarev, A. Packard, W. A. Lester Jr, and M. Frenklach, "Interval prediction of molecular properties in parametrized quantum chemistry," *Physical review letters*, vol. 112, no. 25, p. 253003, 2014.

[218] M. Frenklach, A. Packard, and R. Feeley, "Optimization of reaction models with solution mapping," *Comprehensive Chemical Kinetics*, vol. 42, pp. 243–291, 2007.

[219] X. You, T. Russi, A. Packard, and M. Frenklach, "Optimization of combustion kinetic models on a feasible set," *Proceedings of the Combustion Institute*, vol. 33, no. 1, pp. 509–516, 2011.

[220] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[221] K. Crombecq, D. Gorissen, D. Deschrijver, and T. Dhaene, "A novel hybrid sequential design strategy for global surrogate modeling of computer experiments," *SIAM Journal on Scientific Computing*, vol. 33, no. 4, pp. 1948–1974, 2011.

[222] B. Delaunay *et al.*, "Sur la sphere vide," *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, vol. 7, no. 793-800, pp. 1–2, 1934.

[223] T. Santner, B. Williams, and W. Notz, *The Design and Analysis of Computer Experiments*. Springer, 2003.

[224] W. Welch *et al.*, "Space-filling lhds in computer experiments," in *The Design and Analysis of Computer Experiments*, 1992, p. 134.

[225] M. Handcock, "Cascading lhds," in *The Design and Analysis of Computer Experiments*, 1991, p. 138.

[226] J. Bernardo *et al.*, "Properties of lhds in computer experiments," in *The Design and Analysis of Computer Experiments*, 1992, p. 134.

[227] W. Ji and S. Deng, "Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network," *The Journal of Physical Chemistry A*, vol. 125, no. 4, pp. 1082–1092, 2021.

[228] S. Kim, W. Ji, S. Deng, Y. Ma, and C. Rackauckas, "Stiff neural ordinary differential equations," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 9, 2021.

[229] P. O. Sturm and A. S. Wexler, "Conservation laws in a neural network architecture: enforcing the atom balance of a julia-based photochemical model (v0. 2.0)," *Geoscientific Model Development*, vol. 15, no. 8, pp. 3417–3431, 2022.

[230] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, "Enforcing analytic constraints in neural networks emulating physical systems," *Physical Review Letters*, vol. 126, no. 9, p. 098302, 2021.

[231] Y. Ruckstuhl, T. Janjić, and S. Rasp, "Training a convolutional neural network to conserve mass in data assimilation," *Nonlinear Processes in Geophysics*, vol. 28, no. 1, pp. 111–119, 2021.

[232] A. D. Jagtap, E. Kharazmi, and G. E. Karniadakis, "Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 365, p. 113028, 2020.

[233] F. Djeumou, C. Neary, E. Goubault, S. Putot, and U. Topcu, "Neural networks with physics-informed architectures and constraints for dynamical systems modeling," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 263–277.

[234] H. V. Jagadish, F. Bonchi, T. Eliassi-Rad, L. Getoor, K. Gummadi, and J. Stoyanovich, "The responsibility challenge for data," in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 412–414.

[235] J. Stoyanovich, "Transfat: Translating fairness, accountably and transparency into data science practice," in *1st International Workshop on Processing Information Ethically, PIE@ CAiSE 2019*, 2019.

[236] S. Lebovitz, N. Levina, and H. Lifshitz-Assaf, "Is ai ground truth really 'true'? the dangers of training and evaluating ai tools based on experts' know-what," *The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What (May 4, 2021). Citation: Lebovitz, S., Levina, N., Lifshitz-Assaf, H*, pp. 1501–1525, 2021.

[237] J. S. Saltz and N. Dewar, "Data science ethical considerations: a systematic literature review and proposed project framework," *Ethics and Information Technology*, vol. 21, pp. 197–208, 2019.

[238] S. Barocas and D. Boyd, "Engaging the ethics of data science in practice," *Communications of the ACM*, vol. 60, no. 11, pp. 23–25, 2017.

[239] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[240] H. Werthner, A. Stanger, V. Schiaffonati, P. Knees, L. Hardman, and C. Ghezzi, "Digital humanism: The time is now," *Computer*, vol. 56, no. 1, pp. 138–142, 2023.

[241] D. Firmani, L. Tanca, and R. Torlone, "Ethical dimensions for data quality," *Journal of Data and Information Quality (JDIQ)*, vol. 12, no. 1, pp. 1–5, 2019.

[242] A. Hesse, L. Glenna, C. Hinrichs, R. Chiles, and C. Sachs, "Qualitative research ethics in the big data era," *American Behavioral Scientist*, vol. 63, no. 5, pp. 560–583, 2019.

[243] J. Stoyanovich, S. Abiteboul, and G. Miklau, "Data, responsibly: Fairness, neutrality and transparency in data analysis," in *International Conference on Extending Database Technology*, 2016.

[244] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Communications of the ACM*, vol. 41, no. 2, pp. 79–82, 1998.

[245] D. Jones and B. Simons, *Broken ballots: Will your vote count?* CSLI publications Stanford, 2012.

[246] R. S. Mans, W. M. van der Aalst, R. J. Vanwersch, R. S. Mans, W. M. van der Aalst, and R. J. Vanwersch, "Data quality issues," *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*, pp. 79–88, 2015.

[247] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics*, vol. 154, pp. 72–80, 2014.

[248] J. Du and L. Zhou, "Improving financial data quality using ontologies," *Decision Support Systems*, vol. 54, no. 1, pp. 76–86, 2012.

[249] C. Batini and M. Scannapieco, "Data and information quality: concepts, methodologies and techniques," *Cham: Springer International Publishing*, 2016.

[250] A. Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality," *Journal of Industrial Engineering and Management (JIEM)*, vol. 4, no. 2, pp. 168–193, 2011.

[251] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[252] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of data and analytics*. Auerbach Publications, 2018, pp. 296–299.

[253] S. K. Glaberson, "Coding over the cracks: Predictive analytics and child protection," *Fordham Urb. LJ*, vol. 46, p. 307, 2019.

[254] M. O. Prates, P. H. Avelar, and L. C. Lamb, "Assessing gender bias in machine translation: a case study with google translate," *Neural Computing and Applications*, vol. 32, pp. 6363–6381, 2020.

[255] A. Lillywhite and G. Wolbring, "Coverage of ethics within the artificial intelligence and machine learning academic literature: The case of disabled people," *Assistive Technology*, 2019.

[256] M. Turilli and L. Floridi, "The ethics of information transparency," *Ethics and Information Technology*, vol. 11, pp. 105–112, 2009.

[257] K. Werder, B. Ramesh, and R. Zhang, "Establishing data provenance for responsible artificial intelligence systems," *ACM Transactions on Management Information Systems (TMIS)*, vol. 13, no. 2, pp. 1–23, 2022.

[258] J. A. Tullis and B. Kar, "Where is the provenance? ethical replicability and reproducibility in giscience and its critical applications," *Annals of the American Association of Geographers*, vol. 111, no. 5, pp. 1318–1328, 2021.

[259] D. J. Hand, "Aspects of data ethics in a changing world: Where are we now?" *Big data*, vol. 6, no. 3, pp. 176–190, 2018.

[260] A. Chapman, P. Missier, G. Simonelli, and R. Torlone, "Capturing and querying fine-grained provenance of preprocessing pipelines in data science," *Proceedings of the VLDB Endowment*, vol. 14, no. 4, pp. 507–520, 2020.