Published in Next-Gen Digital Services. A Retrospective and Roadmap for Service Computing of the Future, pp 236-248, LNCS, Springer, 2021

About the Quality of Data and Services in Natural Sciences

Barbara Pernici[0000−0002−2034−9774],
Francesca Ratti[0000−0002−9639−8089], and
Gabriele Scalia[0000−0003−3305−9220]

Department of Electronics, Information and Bioengineering
Politecnico di Milano, Milano, Italy
{barbara.pernici,francesca.ratti,gabriele.scalia}@polimi.it

Abstract. Managing data related to natural sciences poses new and challenging problems as it is impossible to represent reality on a one-to-one scale, and imprecision has to be taken into account, both in data memorization and in its processing. Machine learning has been a key enabler in the context of information extraction from natural sciences data. However, data-driven results are strongly affected by the volume, the sparsity and different types of imprecision in the available sources. Therefore, it becomes pivotal to associate both to data and to data-driven services information about their quality, in order to effectively interpret the results. Different levels of granularity and multiple data modalities captured from the same processes could coexist, due to technological constraints or other intrinsic limiting factors. In addition, different levels of granularity might be also the result of application requirements, and outcomes at multiple levels of precision needs to be provided. Affinities of quality issues in domains such as chemistry, biology, and geoinformatics are discussed in the paper.

Keywords: Data quality · Quality of service · Machine learning.

1 Introduction

The online provisioning of services and data has been studied since the early 2000's in terms of ensuring the characteristics of the services being provided. In particular, quality of service (QoS) characteristics and the problem of representing the quality requirements have been studied in detail [29]. In many cases the provided services are a composition of several services and this poses additional challenges to providers, as the service composition has to perform according to agreed QoS levels.

In his research work, Mike Papazoglou has always promoted the idea of focusing on service management at different levels in service-oriented computing [34, 35]. Furthermore, he pointed out the importance of an infrastructure support for data and process integration. The service execution environment, its underlying infrastructure and the data provided in service requests have very variable characteristics. For this reason, several approaches have been proposed for service management, for providing adaptivity in services and for their compositions in order to guarantee the QoS requirements [32]. Adaptation has been investigated in many papers by Papazoglou towards guaranteeing compliance and QoS with Service Level Agreements (SLA) and contracts, focusing also on the evolution of services (e.g., [2]).

Nowadays, services have become a common infrastructure in many application domains. Nevertheless, the Service Computing manifesto of 2017 [9], advocates that the efforts of researchers have been mainly focused on technological aspects of services, in particular on web services. However, new challenges are posed by emerging technologies. A stronger need of inter-organizational cooperation is growing. The ability of supporting the collection of sensing data from pervasive sensing devices and (re)using data collected by other organizations is required. It is also important to support the human interpretation of the results obtained by services. In a recent Dagstuhl seminar [13] the importance of creating ecosystems for sharing data and providing services in an inter-organizational context was discussed, where the quality of data and services are a key issue.

The goal of this paper is to revisit the research challenges posed by data and services, in the light of requirements emerging in the context of natural sciences and scientific data. In this context, data are characterized by an intrinsic and heterogeneous level of imprecision. In addition, pipelines of data analysis services, which include modeling services, simulation tools and other data-driven services often based on Machine Learning (ML) and Deep Neural Networks (DNN), have added new sources of uncertainties in the process. We will discuss how the service computing principle can be tailored to this context and how imprecision can be managed following QoS principles. Some examples will be derived from chemistry, biology, and geographical spatio-temporal representations.

The paper is structured as follows. In Section 2, we go over related work on adaptivity and compliance and on data representation and imprecision. In Section 3, we discuss challenges and open problems in representing and managing imprecise data in natural sciences. Finally, in Section 4 we discuss how imprecision can be managed within a service computing approach.

2 Related work

In this section, we briefly examine some of the key papers that put the basis for the discussion about providing services in the natural science domain, with a focus on imprecision in data and services.

One of the key issues is to represent and guarantee data quality. The characteristics of data quality modeling and representation for service computing are described in detail in [29], which compares the many dimensions, models, and methods proposed in the area. As discussed in [7], data and information quality has to be managed from a number of points of view. Some of the main issues are related to the structure of information and the representation of data values.

As mentioned in the introduction, adaptivity is advocated as a key feature in composed services. To the purpose of this paper, we focus mainly on quality issues. In the direction of representing QoS in variable contexts, in [3] the issue of varying soft and hard constraints within a contract is analyzed. In [2] the evolution of services, also due to QoS-level induced service changes, is studied, distinguishing between shallow changes, limited to the single service, and deep changes, that have an impact also on other services and providers.

Since the publication of [33], the semantics of conceptual schemas has taken a key role in database design. Different structures can be defined for the same data domain and a systematic approach is needed to identify objects and their relationships in a schema, as well as similar objects in different schemas. Based on [33], some systematic schema similarity assessment metrics have been defined and further developed in [14]. As natural science data are likely to originate from different heterogeneous sources, similar problems can be encountered in integrating them. As discussed in [13], different data interoperability architectures can be envisioned for data integration, data exchange, data repositories, and collaborative data sharing. The semantic associated to the schemas is a critical issue in data ecosystems [13]. In some cases inconsistencies and incompabilities can not be avoided, thus services are needed to access and manage them.

This requires adequate metadata, but "adequate" has a different meaning when the goal changes [26]. Context-aware data quality management supported by an effective metadata management has been recently discussed in [5]. General challenges related to the management of data and data-driven services in the context of scientific data frameworks have been discussed in [38]. In this context, metadata management has a key role.

Another issue impacting data quality and, as a consequence, the quality of services, is that data can be not only imprecise, but also represented at different levels of granularity. Spatio-temporal data is characterized by an intrinsic imprecision and by (sometimes implicit) relationships. The modeling of spatio-temporal data has been studied in [28], integrating work from temporal database literature and spatial data management. Implicit temporal information which can be extracted from temporal relations between events (e.g., before or after) and temporal indeterminacy are difficult to represent and query in conventional databases [11, 4]. Challenges in this directions permeate the analysis of data in natural sciences. The integration of spatial information at different levels of granularity and with a varying accuracy has been recently highlighted as a key issue in various domain, from the integration of single cell biological data [30] to geographical data extracted from social media [25].

3 Representing imprecise data and services

In the following, several open challenges are introduced and discussed. These are characterized by being common to virtually all fields within natural sciences, and derive from data and service quality limitations, in particular related to uncertainty of data and data-driven models. After presenting an overview of the main challenges, relevant directions investigated in recent years are discussed, focusing, in particular, on machine learning-based services. We conclude the section with a discussion on uncertainty of data and services in relation to quality for data-driven applications.

3.1 Overview and challenges

Some recurrent themes characterizing scientific data are the varying levels of granularity, the highly variable quality of the information available (which includes the uncertainty about and originating from the data) and the high number of dimensions (which includes multiple modalities) captured. In turn, the latter can be highly heterogeneous across data points. These characteristics of the data can stem from the model(s) designed to explain them, be the result of technological limitations and of other trade-offs, or be the consequence of a limited knowledge when interpreting and analyzing the gathered information.

The varying granularity, quality and number of collected features characterize the integration and the analysis of scientific experiments from multiple sources and collected over extended periods of times (for example, in the combustion kinetics domain [38]). In this case, technological limitations constrain the *resolution* and the *confidence* of the collected measurements. On top of this, arbitrary choices on the *aggregation level* at which the collected data are described (for example, in scientific papers or repositories) further increase the variability across all directions. Finally, *ambiguities* related to the experimental description, which can also be promoted by flexible and unstructured formats, can further exacerbate these phenomena.

Trade-offs related to the obtainment of the data are often responsible for the variability of its features, with less accurate methods being more cost and time effective. For example, in an ideal world thermochemical properties for all the relevant chemical species would be obtained experimentally or by using high-quality quantum mechanical calculations. However, the cost and time involved would be unbearable, and several other progressively less accurate but more scalable approaches have been proposed, and are used to build varying quality datasets upon which state-of-the-art ML models are trained [22].

A notable example of varying granularity is the *spatial granularity*, observed across many domains beyond that of geographic information analysis. Even though in theory geographic data can be represented at an arbitrary resolution, in practice their extraction in limited-knowledge contexts drastically hinders the available resolution. This is, for example, the case for locations extracted from social media analysis [25]. Here, the location extraction step usually involves data-driven *disambiguation* or crowd-sourcing based activities, which also introduce a varying accuracy in the results. This means that, for example, the location associated to an image extracted from social media is known only up to a certain administrative level (e.g., city or country) even though its "true" location is, in principle, a point in space. Moreover, the inferred location could be only partially correct, for example being correct up to a certain administrative level. Similar issues are central to many other scientific domains. For example, in biology, recent efforts towards the creation of single-cell resolution atlases of organs (e.g., [16]) have highlighted challenges related to the integration of spatial data with a varying resolution, quality and number of modalities [30]. In this case, variability mainly stems from the existence of different technologies with different trade-offs in terms of resolution, throughput, confidence, etc. This has led to the development of integration services to overcome the limitations of each individual technology, and to enable analyses at multiple resolution scales.

Even though only based on few examples, the above discussion makes evident how many recurrent themes can benefit from the investigation of common solutions.

3.2 Requirements and application solutions

The recent focus on the development of data-driven scientific frameworks, atlases and computing pipelines has highlighted a set of specific requirements and architectural needs to support them. While often arising in a specific domain, these requirements are for the majority general, and shared among scientific fields.

Recently, a set of requirements to enhance the capabilities of data-driven scientific frameworks has been discussed in [38]. These include:

- The continuous and semantic *multi-source integration*, with a focus on the heterogeneous quality of the sources and their mutual dependencies.
- The *dynamic acquisition* of new information ("open-world" assumption).
- the *continuous dynamic validation* of stored information as new knowledge and sources are acquired, accounting for data and model uncertainties.

Though not exhaustive, these requirements provide an abstract framework to describe the application solutions recently investigated across domains. In the following, these requirements are generalized, and key application solutions are framed within them.

Multi-source Integration. As previously mentioned, the information conveyed by the different data sources is heterogeneous and can vary largely in terms of resolution, accuracy and coverage. For its importance, integration is a prerequisite in most of the other activities [30]. One additional challenge often faced in this area is the lack of references or ontologies to drive the integration [38]. Instead, little or no prior information is often available, and complementary strengths of the different sources need to be exploited to automatically generate ontologies or achieve integration in the absence of curated references. Focusing on data-driven methods, we can identify several directions.

Different sources can be individually analyzed and their outputs jointly used to overcome the limitations of each individual source. This is often the case when the analysis of one source can enhance the information extraction process from the others (see also *dynamic acquisition*). For example, it has been shown how the integration of spatial proteomics data and protein-protein interaction network data enables the extraction of more information and increases the predictive power [39]. On the same line, extracting geographical information from multiple social media, an iterative triangulation-based approach can overcome the limitations of each individual source in terms of accuracy, volume and available modalities [6].

In many cases, different sources are *fused* and *aligned* in a common shared space. This unsupervised process is particularly challenging in the absence of reference data. Finding common sources of variation in heterogeneous data is key in health research, metabolomics, epigenetics and epidemiology, to name a few [23]. It has been recently shown how, by detecting common sources of variation, single-cell transcriptomic data can be effectively integrated across different conditions, technologies, species, and modalities [12]. Unsupervised deep learning methods have been used to achieve a similar goal [19]. In all cases, the aim is to derive a shared manifold across data features. Even though promising, existing methods are characterized by scalability and flexibility limitations [12, 19].

Finally, a class of integration strategies characterizing ML-based services is based on *domain adaptation.* Strategies such as *transfer learning* (transferring the knowledge learned in a source domain to a target domain) and multi-task learning (using multi-task objectives to implicitly learn and exploit a shared latent space) have been recently used to cope with heterogeneous sources in natural sciences domains. Transfer learning techniques are particularly effective in these domains, given the challenges usually faced constructing large-scale well-annotated datasets [40]. Transfer learning has shown promising results in molecular property prediction, integrating small sets of high-accuracy data to larger set of less accurate datasets [22]. Similar strategies have been also used integrating single cell transcriptomics across batches and datasets [43, 42].

Dynamic Acquisition. Even though including new samples usually improves the performance of data-driven algorithms (assuming new data has comparable precision), the extent of this effect largely depends on *which* new samples are available. By making an "open-world" assumption, a system accounts for the existence of data samples external to it, which can be queried/produced (with an associate cost) and integrated into the system (e.g., in the training phase). In this setting, the best predicate for querying new information depends, in general, on the already available information, the cost associated to gather the new information and on a background knowledge [38]. The net result of this approach is a virtuous refinement cycle.

This cycle can follow also the time direction. For example, as time passes, new information can be made available and already collected data can drive more accurate queries. This approach has been used, for example, to extract geographical information from social media and to iteratively evaluate the relevance of collected contents to refine the search keywords [6].

In the ML community, the iterative refinement of a model through the acquisition of new training data is named *active learning.* The acquisition of new training samples usually involves time consuming and/or costly operations (e.g., human in the loop), thus the need to optimize new queries. This is often the case in natural sciences, where DNN-based active learning frameworks have been recently proposed to query manual annotations [21] or other more accurate models

[31]. An active learning framework often includes the calculation of the *uncertainty* of the model over the predictions, which contributes to the selection of the best new samples to be added.

Continuous Dynamic Validation. In a data-driven framework, data validation is both a prerequisite for further analyses and the result of data integration and cleaning services. These two activities can follow a virtuous cycle.

The validation of input data to ML pipelines is subject of active research and has to be tackled from different perspectives. For example, [10] distinguishes between *single-batch*, where the focus is on highlighting anomalies in a single batch of data, *inter-batch*, to capture significant changes between training and serving data or different batches of training data, and *model testing*, to ensure that there are no assumptions in the training code that are not reflected in the data.

When little or no prior references/ontologies are available, data validation and curation activities can follow integration and acquisition steps. Comparisons across data features and modalities, possibly with additional data gathered through targeted queries, can enable the validation of existing data in the absence of prior ground truth. For example, a transfer learning approach has been used to integrate and correct multiple RNA sequencing batches [43] and to improve the quality of noisy and sparse single-cell transcriptomics data [42]. Cross-comparison of experimental datasets and models extracted from the literature can help discovering inconsistencies [24, 38].

The goal of the above discussion is to link general data and service quality management requirements to recent data-driven application solutions explored across domains. The similarities pointed out should drive the research of both general techniques and shared theoretical frameworks. One key feature underlying all the discussed requirements is the management of *uncertainty* in the data and introduced by services. This becomes particularly important when data are inherently noisy and for machine-learning based services. This is discussed in the following.

3.3 Uncertainty of data and services

The quality of the data is central to the definition of *value* in services. Indeed, the value obtainable from service orchestration hinges on the quality of the data exchanged among the orchestrated services [1]. Being at the highest level in the computing value chain, the quality of a service is generally affected by the knowledge it relies on, which ultimately depends on the underlying data. This relationship, which usually exists in terms of data exchanged between services, takes a much wider significance for data-driven and, especially, ML-based services. In the latter case, indeed, available data contribute to the definition of functions processing new data (think, for example, of a service based on a trained ML models, which output/quality depends on the underlying dataset and its quality). Therefore, in this context, the relationship between data and service quality needs to be revised [8].

We observe how progresses in the deep learning community have recently led to the development of models that can efficiently compute calibrated uncertainties over their predictions. Notably, approximate Bayesian DNNs have been proposed as principled methods to separately compute the *epistemic uncertainty*, which stems from the model's ignorance about the underlying model (e.g., insufficient training data) and the *aleatoric uncertainty*, which intrinsically characterizes the data (e.g., experimental noise, stochasticity, etc.) [18, 27]. The recent spread of these models in fields such as chemistry, biology and medicine [37, 17, 41] has shown promise in explaining and discerning a model's predictions when trained on noisy, incomplete and heterogeneous data. However, some challenges related to the robustness and the interpretability of the estimated uncertainties still remain [37].

Another peculiar feature of ML-based services is the relationship between data volume and uncertainty in the results. Epistemic uncertainty can be explained away given enough data. For this reason, expanding the dataset, even through the usage of *data augmentation* techniques, can enhance the accuracy of the trained model and, consequently, the quality of the resulting service. In this case, the optimal accuracy consists in the right balance between volume and quality of the underlying dataset (with data augmentation techniques promoting the first while, potentially, hindering the second). Outcomes of ML models can complement and augment experimental datasets (with, in particular, DNN-based models being particularly effective to approximate complex natural phenomena), thus ultimately increasing the quality of resulting data-driven services [15, 36].

The above discussion highlights some interdisciplinary research directions which should be investigated in the future:

- How to effectively *transfer* quality dimensions (including uncertainty) back and forth between services and data. Indeed, in this context, data define and refine the services through ML techniques, while, at the same time, datasets are enhanced and extended by other services (e.g., active learning, data cleaning and data augmentation routines).
- How to assess and store the *evolving quality* of the data, distinguishing between "inherent" data quality and other quality indicators progressively introduced by the analysis models, which also take into account the relationship with the (evolving) quality of other data and services.
- How to integrate ML *explainability* techniques [20, 8] to data and service quality management routines and orchestration.

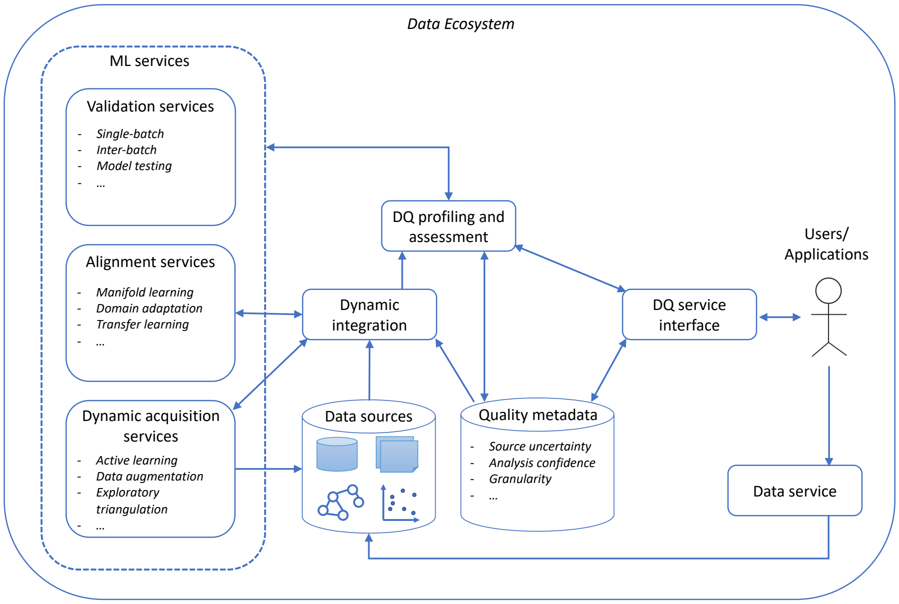

Fig. 1. Schematic architecture of the framework

4 Managing imprecision with services

After the discussion on the origin and the characteristics of imprecision in data and service in the context of natural science at large presented in Section 3, in this section we propose a general architecture that highlights how a contract-based and adaptive approach can support the requirements previously discussed.

The proposed architecture in Figure 1 generalizes the one presented in [38], and contextualizes the machine learning services presented in Section 3 within the identified requirements. At the same time, the architecture extends the Data Quality Service Architecture introduced in [5], with metadata management (which, in our discussion, is mainly represented by *uncertainty management*) having a key role.

While the *dynamic integration* component handles the integration of different sources and metadata at the format/schema level, specific *alignment services* handle integration at the conceptual level. All the ML services interact with and enrich the *quality metadata* through the core *DQ profiling and assessment* service. The *dynamic acquisition services* also feed the data sources, thus allowing an iterative process. Other than feeding new data to the framework, the user/application interacts with the system through the *DQ service interface*. Through this, it has access to a complete quality overview.

5 Concluding remarks

In this paper we discussed how a service computing approach can be tailored to the needs of data management in natural sciences. Focusing on the requirements emerging in this context from data and data-driven services (in particular, ML services), we discussed quality-related challenges and application directions, paying attention in particular on uncertainty management. We proposed a general architecture highlighting the advantages of a contract-based and adaptive approach, taking QoS constraints into consideration.

This work has highlighted several directions which necessitate further investigation. First of all, uncertainty estimation needs to become a central part of scientific data ecosystems. In this respect, investigating how to effectively transfer uncertainty properties back and forth between services and data, taking into account the evolving nature of both, represents an open challenge. On top of this, ML explainability techniques should be integrated to data and service quality management, ideally enriching the DQ service interface presented to the Users/Applications. In addition, if an adaptive approach is pursued, the stability of results should be evaluated, to assess the impact of adaptations in the ecosystem. Finally, as illustrated in [9], data acquisition, integration and validation services, could benefit from crowdsourcing activities. This approach is currently being proposed in the Crowd4SDG project[1], where citizen science is going to be supported by decision making/collaborative platforms and crowdsourcing tools.

Acknowledgements

This work was funded by the European Commission H2020 project Crowd4SDG "Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience" under project No. 872944. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

[1] http://www.crowd4sdg.eu/

References

1. Ameller, D., Illa, X.B., Collell, O., Costal, D., Franch, X., Papazoglou, M.P.: Development of service-oriented architectures using model-driven development: A mapping study. Inf. Softw. Technol. **62**, 42–66 (2015). https://doi.org/10.1016/j.infsof.2015.02.006, https://doi.org/10.1016/j.infsof.2015.02.006
2. Andrikopoulos, V., Benbernou, S., Papazoglou, M.P.: On the evolution of services. IEEE Trans. Software Eng. **38**(3), 609–628 (2012). https://doi.org/10.1109/TSE.2011.22, https://doi.org/10.1109/TSE.2011.22
3. Andrikopoulos, V., Fugini, M., Papazoglou, M.P., Parkin, M., Pernici, B., Siadat, S.H.: QoS contract formation and evolution. In: Buccafurri, F., Semeraro, G. (eds.) E-Commerce and Web Technologies, 11th International Conference, EC-Web 2010, Bilbao, Spain, September 1-3, 2010. Proceedings. Lecture Notes in Business Information Processing, vol. 61, pp. 119–130. Springer (2010). https://doi.org/10.1007/978-3-642-15208-5_11, https://doi.org/10.1007/978-3-642-15208-5_11
4. Anselma, L., Piovesan, L., Terenziani, P.: Dealing with temporal indeterminacy in relational databases: An AI methodology. AI Commun. **32**(3), 207–221 (2019). https://doi.org/10.3233/AIC-190619, https://doi.org/10.3233/AIC-190619
5. Ardagna, D., Cappiello, C., Samá, W., Vitali, M.: Context-aware data quality assessment for big data. Future Gener. Comput. Syst. **89**, 548–562 (2018). https://doi.org/10.1016/j.future.2018.07.014, https://doi.org/10.1016/j.future.2018.07.014
6. Autelitano, A., Pernici, B., Scalia, G.: Spatio-temporal mining of keywords for social media cross-social crawling of emergency events. Geoinformatica **23**(3), 425–447 (2019)
7. Batini, C., Scannapieco, M.: Data and Information Quality - Dimensions, Principles and Techniques. Data-Centric Systems and Applications, Springer (2016). https://doi.org/10.1007/978-3-319-24106-7, https://doi.org/10.1007/978-3-319-24106-7
8. Bertossi, L., Geerts, F.: Data quality and explainable AI. Journal of Data and Information Quality (JDIQ) **12**(2), 1–9 (2020)
9. Bouguettaya, A., Singh, M.P., Huhns, M.N., Sheng, Q.Z., Dong, H., Yu, Q., Neiat, A.G., Mistry, S., Benatallah, B., Medjahed, B., Ouzzani, M., Casati, F., Liu, X., Wang, H., Georgakopoulos, D., Chen, L., Nepal, S., Malik, Z., Erradi, A., Wang, Y., Blake, M.B., Dustdar, S., Leymann, F., Papazoglou, M.P.: A service computing manifesto: The next 10 years. Commun. ACM **60**(4), 64–72 (2017). https://doi.org/10.1145/2983528, https://doi.org/10.1145/2983528
10. Breck, E., Polyzotis, N., Roy, S., Whang, S., Zinkevich, M.: Data validation for machine learning. In: Talwalkar, A., Smith, V., Zaharia, M. (eds.) Proceedings of Machine Learning and Systems 2019, MLSys 2019, Stanford, CA, USA, March 31 - April 2, 2019. mlsys.org (2019), https://proceedings.mlsys.org/book/267.pdf
11. Brusoni, V., Console, L., Terenziani, P., Pernici, B.: Qualitative and quantitative temporal constraints and relational databases: Theory, architecture, and applications. IEEE Trans. Knowl. Data Eng. **11**(6), 948–968 (1999). https://doi.org/10.1109/69.824613, https://doi.org/10.1109/69.824613
12. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology **36**(5), 411–420 (2018)

13. Cappiello, C., Gal, A., Jarke, M., Rehof, J.: Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391). Dagstuhl Reports **9**(9), 66–134 (2020). https://doi.org/10.4230/DagRep.9.9.66, https://drops.dagstuhl.de/opus/volltexte/2020/11845
14. Castano, S., De Antonellis, V., Fugini, M.G., Pernici, B.: Conceptual schema analysis: Techniques and applications. ACM Trans. Database Syst. **23**(3), 286–332 (1998). https://doi.org/10.1145/293910.293150, https://doi.org/10.1145/293910.293150
15. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al.: Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface **15**(141), 20170387 (2018)
16. Consortium, H., et al.: The human body at cellular resolution: The NIH Human Biomolecular Atlas Program. Nature **574**(7777), 187 (2019)
17. Fauw, J.D., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C.O., Raine, R., Hughes, J., Sim, D.A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P.T., Suleyman, M., Cornebise, J., Keane, P.A., Ronneberger, O.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. **24**(9), 1342–1350 (2018), http://lmb.informatik.uni-freiburg.de/Publications/2018/Ron18
18. Fox, C.R., Ülkümen, G.: Distinguishing two dimensions of uncertainty, vol. 14, chap. 1. Universitetsforlaget Oslo (2011)
19. Gala, R., Gouwens, N., Yao, Z., Budzillo, A., Penn, O., Tasic, B., Murphy, G., Zeng, H., Sümbül, U.: A coupled autoencoder approach for multi-modal analysis of cell types. In: Advances in Neural Information Processing Systems. pp. 9267–9276 (2019)
20. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89. IEEE (2018)
21. Gilyazev, R., Turdakov, D.Y.: Active learning and crowdsourcing: A survey of optimization methods for data labeling. Programming and Computer Software **44**(6), 476–491 (2018)
22. Grambow, C.A., Li, Y.P., Green, W.H.: Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. The Journal of Physical Chemistry A **123**(27), 5826–5835 (2019)
23. Gu, Z., de Schipper, N.C., Van Deun, K.: Variable selection in the regularized simultaneous component analysis method for multi-source data integration. Scientific Reports **9**(1), 1–21 (2019)
24. Hansen, N., He, X., Griggs, R., Moshammer, K.: Knowledge generation through data research: New validation targets for the refinement of kinetic mechanisms. Proceedings of the Combustion Institute (2018)
25. Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J.L., Van Achte, T., Zeug, G., Mondardini, M.R.R., Grandoni, D., et al.: E2mC: improving emergency management service practice through social media and crowdsourcing analysis in near real time. Sensors **17**(12), 2766 (2017)
26. Jagadish, H.: Big data and science: Myths and reality. Big Data Research **2**(2), 49–52 (2015)

27. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5580–5590. NIPS'17 (2017), http://dl.acm.org/citation.cfm?id=3295222.3295309
28. Koubarakis, M., Sellis, T.K., Frank, A.U., Grumbach, S., Güting, R.H., Jensen, C.S., Lorentzos, N.A., Manolopoulos, Y., Nardelli, E., Pernici, B., Schek, H., Scholl, M., Theodoulidis, B., Tryfona, N. (eds.): Spatio-Temporal Databases: The CHOROCHRONOS Approach, Lecture Notes in Computer Science, vol. 2520. Springer (2003). https://doi.org/10.1007/b83622, https://doi.org/10.1007/b83622
29. Kritikos, K., Pernici, B., Plebani, P., Cappiello, C., Comuzzi, M., Benbernou, S., Brandic, I., Kertész, A., Parkin, M., Carro, M.: A survey on service quality description. ACM Comput. Surv. **46**(1), 1:1–1:58 (2013). https://doi.org/10.1145/2522968.2522969, https://doi.org/10.1145/2522968.2522969
30. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al.: Eleven grand challenges in single-cell data science. Genome Biology **21**(1), 1–35 (2020)
31. Li, Y.P., Han, K., Grambow, C.A., Green, W.H.: Self-evolving machine: A continuously improving model for molecular thermochemistry. The Journal of Physical Chemistry A **123**(10), 2142–2152 (2019)
32. Metzger, A., Pohl, K., Papazoglou, M.P., Di Nitto, E., Marconi, A., Karastoyanova, D.: Research challenges on adaptive software and services in the future internet: towards an S-Cube research roadmap. In: Metzger, A., Pohl, K., Papazoglou, M.P. (eds.) First International Workshop on European Software Services and Systems Research - Results and Challenges, S-Cube 2012, Zurich, Switzerland, June 5, 2012. pp. 1–7. IEEE (2012). https://doi.org/10.1109/S-Cube.2012.6225501, https://doi.org/10.1109/S-Cube.2012.6225501
33. Papazoglou, M.P.: Unraveling the semantics of conceptual schemas. Commun. ACM **38**(9), 80–94 (1995). https://doi.org/10.1145/223248.223275, https://doi.org/10.1145/223248.223275
34. Papazoglou, M.P., Georgakopoulos, D.: Introduction. Commun. ACM **46**(10), 24–28 (2003). https://doi.org/10.1145/944217.944233, https://doi.org/10.1145/944217.944233
35. Papazoglou, M.P., Traverso, P., Dustdar, S., Leymann, F.: Service-oriented computing: State of the art and research challenges. IEEE Computer **40**(11), 38–45 (2007). https://doi.org/10.1109/MC.2007.400, https://doi.org/10.1109/MC.2007.400
36. Ratti, F., Scalia, G., Pernici, B., Magarini, M.: A data-driven approach to optimize bounds on the capacity of the molecular channel. In: Accepted in 2020 IEEE Global Communications Conference (GLOBECOM). IEEE (2020)
37. Scalia, G., Grambow, C.A., Pernici, B., Li, Y.P., Green, W.H.: Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. Journal of Chemical Information and Modeling **60**(6), 2697–2717 (2020). https://doi.org/10.1021/acs.jcim.9b00975
38. Scalia, G., Pelucchi, M., Stagni, A., Cuoci, A., Faravelli, T., Pernici, B.: Towards a scientific data framework to support scientific model development. Data Science **2**(1-2), 245–273 (2019)
39. Squires, S., Ewing, R., Prügel-Bennett, A., Niranjan, M.: A method of integrating spatial proteomics and protein-protein interaction network data. In: International Conference on Neural Information Processing. pp. 782–790. Springer (2017)

40. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. pp. 270–279. Springer (2018)
41. Tomašev, N., Glorot, X., Rae, J., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C., Peterson, K., Reeves, R., Hassabis, D., King, D., Suleyman, M., Back, T., Nielson, C., Ledsam, J., Mohamed, S.: A clinically applicable approach to continuous prediction of future acute kidney injury. Nature **572**(7767), 116–119 (2019). https://doi.org/10.1038/s41586-019-1390-1
42. Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., Zhang, N.R.: Data denoising with transfer learning in single-cell transcriptomics. Nature Methods **16**(9), 875–878 (2019)
43. Wang, T., Johnson, T.S., Shao, W., Lu, Z., Helm, B.R., Zhang, J., Huang, K.: BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biology **20**(1), 1–15 (2019)