Enabling data spaces: Existing developments and challenges

Gürkan Solmaz, Flavio Cirillo*, Jonathan Fürst†, Tobias Jacobs, Martin Bauer, Ernö Kovacs*, Juan Ramón Santana, Luis Sánchez‡
NEC Laboratories Europe, Germany*, Zurich Univ. of Applied Sciences, Switzerland†,Universidad de Cantabria, Spain‡
firstname.surname@neclab.eu*, jonathan.fuerst@zhaw.ch†, jrsantana@tlmat.unican.es, lsanchez@tlmat.unican.es‡

ABSTRACT

This paper focuses on the concept of *data spaces*, which can serve as a basis for the future data economy. In data spaces, applicable to various business domains, stakeholders will be able to share data with each other in a controlled way. First, the paper describes the real motivations and needs for enabling data spaces. Second, it highlights the major technical developments in the area of data spaces in the light of open ecosystems and standards. Lastly, it focuses on two key challenges for enabling data spaces: 1) Data interoperability, 2) Data value generation. As a concrete data spaces solution example, this paper proposes the *"Green Twin"* use case that can be developed as a carbon neutrality solution in the domains of mobility and smart cities.

CCS CONCEPTS

• **Computer systems organization** → *Distributed architectures*; • **General and reference** → Design; • **Information systems** → **Data management systems**.

KEYWORDS

data spaces, Gaia-X, IDSA, FIWARE, data analytics, digital twins

ACM Reference Format:
Gürkan Solmaz, Flavio Cirillo*, Jonathan Fürst†, Tobias Jacobs, Martin Bauer, Ernö Kovacs*,, Juan Ramón Santana, Luis Sánchez‡. 2022. Enabling data spaces: Existing developments and challenges. In *Data Economy (DE '22), December 9, 2022, Roma, Italy.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3565011.3569058

1 INTRODUCTION

Data spaces is a concept that gains increasing attention globally from industries and research communities. This concept serves as an abstraction for data management in case where many stakeholders are involved and exchange data with each other. The easy data exchange between the stakeholders will generate value, especially in combination with data analytics. New trading mechanisms can allow stakeholders to cooperate with each other based on the value of the exchanged data and the analytics services. For instance, in a city, the public transportation company and local businesses might participate in a data space in which businesses benefit from a better retail demand prediction, while the transportation company can optimize traffic management. Thus, data exchanges and analytics in the data spaces would together create data economy in a smart city. The concept of data economy is previously discussed for smart cities [21] from the internet of things perspective.

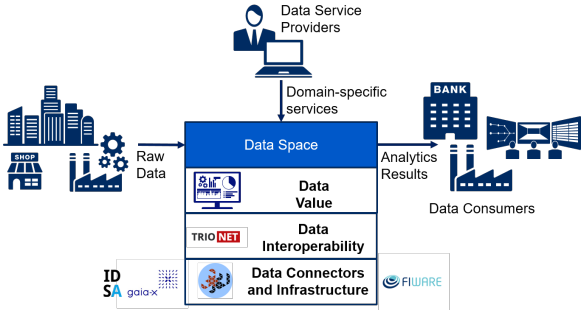

Figure 1: Stakeholders and key layers of the data space.

The efforts toward building data spaces are currently led by Gaia-X [13] and International Data Spaces Association (IDSA) [3]. The synergies between these two communities [1] receive a broad attention in Europe [5, 6] and beyond [18, 20]. The data connectors such as the *IDS Connector* would provide easy data exchanges. Furthermore, standard-based open ecosystems such as FIWARE [8] provide building blocks for data platforms such as data brokering through standardized data model [14].

The stakeholders in a data space may be data providers and/or consumers, as well as service providers. The basic concept is illustrated in Fig. 1. Various data providers from different domains or verticals can share their data within the data space. The data space should be able to manage a plethora of data sources with different data models or representations. Service providers can operate their services by accessing the shared data space. For instance, they can run data analytics services. Lastly, the analytics results from the service providers are shared with the data consumers. The figure includes three layers which are considered as the *key enablers* of the data spaces. IDSA, Gaia-X, and FIWARE currently work on necessary building blocks for the data connectors and infrastructure. The above two layers are to be addressed to provide harmonized and re-usable data for many stakeholders and generate value using the interoperable data.

This paper focuses on the existing developments and open challenges for enabling data spaces to support the future of the data economy. To make the concepts discussed throughout the paper more clear, an example use case, namely the *Green Twin*, is considered. The Green Twin use case regards the goal of carbon neutrality as a global challenge with many stakeholders in the environment. The example use case is described in more detail in Section 2. To

DE '22, December 9, 2022, Roma, Italy

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3565011.3569058

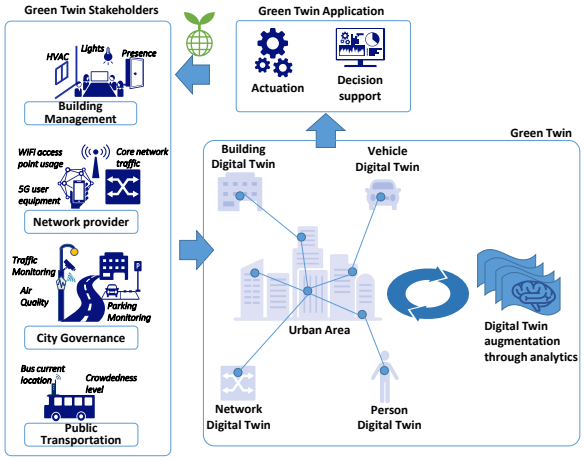

Figure 2: The Green Twin use case for carbon neutrality.

understand the current status of data spaces, Section 3 gives an overview of the recent advancements from a technical point of view in Gaia-X, IDSA, and FIWARE, in particular through necessary standardized building blocks such as data connectors and data brokering. This section includes a technical solution example in terms of the existing data sovereignty and decentralization, data models and semantics.

On top of the data connectors and infrastructure, this paper describes two major challenges: 1) Data interoperability and 2) Data Value generation. Data interoperability considers the interoperability between stakeholders on the higher data layer, as opposed to the communication layer. Section 4.1 describes the data interoperability based on the automated and semi-automated harmonization of the data models. Data value generation considers advanced data processing functions with automated re-use capabilities to generate value in the data space persistently. Section 4.2 describes the data value generation through re-usable machine learning models and easy configuration of prediction models. Both sections include a technical solution example for the Green Twin use case based on the data interoperability and data value aspects. Finally, Section 5 concludes the paper, highlighting the key aspects presented.

2 EXAMPLE USE CASE: GREEN TWIN

As an application example of the data space, the *Green Twin* use case is illustrated in Fig. 2. The use case includes services for monitoring and coordinating infrastructure operations and human activities aiming at reducing energy consumption, while still enhancing quality of life. The Green Twin makes use of four categories of digital twins in urban environments, which are digital representations of real things: Building twin, vehicle twin, person twin and network twin.

The Green Twin consists of static (3D models, HVAC datasheets) and dynamic (smart meters, wearable, in-vehicle sensors) data. Digital twin instances are related to each other through relationship links (e.g., vehicleA is parked underneath buildingB). Data analytics processes, attached to digital twins, work on top of the digital twin data to infer current status, predict future status, simulate hypothetical conditions, and decide upon actions. The outcomes of those processes are part of the digital twins. A digital twin instance is shared by multiple stakeholders, while each of them holds only a partial set of information and analytics. Typical data providers of the Green Twin are system manufacturers (vehicle manufacturers, HVAC manufacturers, building constructors) and facility providers (energy providers, water providers, etc.). Typical data consumers are building management and policymakers. Often, a single stakeholder acts as both provider and consumer. For example, the building management might provide data about the usage of the building and at the same time use the data in the data space to optimize their operations.

The Green Twin use case envisages a complexity of the digitally represented reality. For example, it encompasses multiple buildings, each consisting of multiple rooms, corridors and staircases. Each building is topologically placed in an urban area with certain relations between each other (e.g., a laboratory building is complementary to a nearby lecture building of a university. A similar case are rooms in the same building). Additional examples of sub-systems of the digital twin use case are parking lots, water infrastructure and power grid.

3 EXISTING TECHNICAL DEVELOPMENTS

Data exchange is at the core of the data economy, as the main enabler of the global digital market. Therefore, a trustworthy framework becomes essential to expand and foster a thriving business ecosystem around data, that would nurture the data economy. However, nowadays, the data economy emanates from traditional centralized data storage solutions. As a consequence, it has become a fragmented ecosystem, following closed proprietary solutions that hinder the evolution of new business developments based on data economy, also due to the lack of well-established interoperable open solutions. In fact, the dominant market position of a small set of actors introduces vendor lock-in limitations, forcing users to yield the control of their data, and creates new entry barriers for the smaller ones. Besides, its impact also led to a lack of trust, in which entities cannot exchange data while keeping their control, specially once data access has been granted to data consumers. In this sense, data spaces can reduce such limitations due to their open decentralized approach.

During the last years, several initiatives have appeared that foster the data ecosystem through the provision of novel interoperable open solutions for the creation of data spaces. They enable data producers to keep the control over their data while creating the basis for interoperable data exchange. The goal is that data consumers are able to access different data sources without the limitations stemming from their heterogeneity, either related to their semantics or their technical access requirements.

This section introduces the data space initiatives and the key aspects they address. It describes the existing technical developments in data spaces and presents their adoption required for the realisation of the Green Twin use case.

3.1 Data sovereignty and decentralization

As aforementioned, one of the key issues that hinders the data economy is that data providers typically have to trust big operators

of data infrastructures and do not retain any control over their data, once it has been exposed to a consumer. Thus, many potential data providers are reluctant to share and exchange their data. To address this issue, the concept of *data space* has been introduced. The key idea of a data space is that data providers keep control of their data, i.e. typically store it on their own premises. Potential data consumers directly interact with data providers, negotiating the conditions under which data can be accessed and used. Only once agreement on the conditions has been confirmed, data consumers get access to the data for the agreed use. This *staying in control* of the data is also referred to as *data sovereignty*. As a foundation, it requires a trustworthy framework with clear rules that participants of the data space have to comply with.

The International Data Spaces Association (IDSA) [3] has promoted the concept of data spaces and developed the IDS Reference Architecture Model (IDS-RAM) [2]. IDS-RAM defines the required standards, control and enforcement rules for data exchange among different participants in a data space, specifying their components and mechanisms. The technical definition of the IDS-RAM components and standards is released as part of the International Data Spaces Global (IDS-G) set of specifications [4].

The key element of IDS data spaces is the *IDS connector*. It acts as the entry point to an IDS data space. Any transaction within an IDS data space has to be carried out through a certified IDS Connector, where specific usage control rules can be applied to any type of data. Thus, any data transaction between two parties requires an explicit agreement made through their corresponding IDS Connectors. This way, the restrictions pertaining to the access and usage of data imposed by the data owners can be technically enforced. This is particularly important as traditional systems only consider access control policies, which do not enforce data usage policies after the data have been accessed. Moreover, IDS Connector can also host certified apps, accessible from a central IDS App Store component, increasing the trust on sensitive data exchange.

Furthermore, an IDS data space has the following components:

- *Identity Provider* - named as Dynamic Attribute Provisioning Service (DAPS), which implements OAuth2 authentication.
- *Clearing House* - logs any transaction carried out within the IDS data space, enabling the auditing of data transactions within the data space.
- *Metadata Broker* - stores the metadata related to data providers that belong to the data space, enabling data discovery. The metadata is described based on a common ontology called Information Model [24], which also describes the actors and their interactions within a data space.
- *Vocabulary Hub* - enables the storage of known ontologies that can be linked to describe the data being exchanged.

To enable trust, certification plays a key role in IDS data spaces. In particular, the IDS Connectors have to be certified for the processes of participants related to exchanging and using data.

In 2019, the Gaia-X association was created with the goal of defining a framework with related policies and rules to enable the creation of federation cloud services across cloud-based service providers [16]. As in IDSA, the concepts of data sovereignty and trust are of key importance. Gaia-X not only enables the decentralization of data-related services, but also most infrastructure services, e.g. it builds on the concept of self-sovereign identities based on the W3C Decentralized Identifiers (DIDs) [25].

Furthermore, all services and all participants in the ecosystem have self-descriptions based on verifiable credentials (VC) [26], which can be combined to verifiable presentations (VP). The verifiable credentials are signed by trusted parties attesting the validity of the included claims. To create a trusted environment, there are Gaia-X trust anchors and chains of trust are founded on these. The focus of Gaia-X is on digitizing the description of all aspects required for cloud services and data exchange in particular, standardizing the vocabulary and the required elements of the self-description. The goal is to give back control to the participants and create a more level playing field for cloud service providers. In Gaia-X, the focus is clearly on the meta level, i.e. on describing services, participants and data to be exchanged, whereas the data exchange itself is out-of-scope, e.g. a container as defined by IDSA is considered to be one possible implementation technology.

In short, Gaia-X and IDSA provide rules and framework components for enabling trusted data exchange, but they do not define the data models and the detailed interactions required to achieve interoperable data exchange between participants.

3.2 Interfacing and data modelling

Heterogeneity in the access to data is one of the most important factors that limit the expansion of the data economy, as it hinders the applicability of solutions that would require data coming from different sources. Therefore, data interoperability, both with respect to data access and to data modelling, needs to be guaranteed within data spaces to ease the development of portable and replicable solutions. To achieve data interoperability, the technological interfacing as well as the data modelling employed in the data exchange needs to be agreed.

However, as shown in Section 3.1, the existing initiatives aiming at defining the technical soft infrastructure of future data spaces are not dealing with the data modelling. They restrict their specifications to exchange of metadata related to the transaction, but not to the actual data exchanged, or simply specifies the data representation formats, without defining any recommendation/specification in terms of actual data semantics. Besides, data distribution specifications are not standardized, nor even harmonized, among data providing platforms. Hence, they can use their own proprietary solutions that need to be employed by consumers willing to access data from the same data space but different providers.

In this regard, the Next Generation Service Interfaces Linked Data (NGSI-LD) standard [9] can be highlighted as a candidate to harmonize the specifications of data access and enable the data interoperability among different data providers and consumers within data spaces. NGSI-LD is an ETSI standard that provides a fully-fledged specification to enable context data management. In this sense, NGSI-LD can facilitate the access to context information by defining the Application Programming Interfaces (API) and the data models to be used by the different participants within a data space. This standard is the core interface of the FIWARE open source ecosystem and is already being employed in many real-world pilots [7, 22] providing a flexible and reliable way to overcome the limitations of data interoperability in scenarios where it is necessary

to harmonize the access to data coming from heterogeneous data sources. FIWARE provides a set of open-source components that can be used for building data platforms. A central FIWARE component is the Context Broker [14] implementing the NGSI-LD API.

The NGSI-LD API is based on an abstract information model based on the concept of *entity*, where entities have types, properties and relationships. There are already some existing initiatives that are targeting the definition of a corpus of NGSI-LD compatible data models that can be used as a reference for semantically modelling the data to be exchanged within the future data spaces. Among them, it is worth mentioning the Smart Data Models program [15], which is meant to underpin the semantic interoperability of context information in data spaces. The catalogue of data models that are being created leverages the linked-data nature of NGSI-LD information model and, through the creation or mapping of existing ontologies, provides a common semantic description of terms that can be used by any data provider, ensuring their semantic interoperability.

All in all, these technical developments complement the ecosystem defined by the IDSA and Gaia-X initiatives, providing the required tools for the semantically-enabled data interoperability within data spaces, and fostering the portability of services and applications that make use of shared data. In particular, Gaia-X, IDSA, FIWARE and the Big Data Value Association (BDVA) have created the Data Spaces Business Alliance (DSBA) [11] to converge on a common approach and create building blocks for the Data Economy.

3.3 Technical solution for Green Twin

Existing data space technologies are required to realize the Green Twin use case. Among them, data sovereignty solutions, such as the ones presented in Section 3.1, enable data providers to exchange data while keeping their control. This is particularly important for sensitive information that could result in future business disadvantages. For instance, transportation companies that might want to share fuel consumption information exclusively for research purposes, while limiting their access for other activities that could benefit potential competitors.

Similarly, the use of a common interface and data model would ease the data exchange within the Green Twin use case, enabling different stakeholders to implement their solutions while reducing the development efforts. For instance, a building management service developer could be interested on the combination of data from different providers, such as the network provider, the urban mobility manager or the municipality, to train machine learning models to reduce the building carbon footprint by triggering certain actions. Further, the semantics built on top of NGSI-LD allow the transparent blending of information as a single knowledge graph, and the developer as data consumer could also be subscribed to such information and receive a continuous data flow thanks to the subscription functionalities provided by the FIWARE Context Brokers.

4 OPEN CHALLENGES

As described in the previous sections, there are key technologies in place for building data spaces. However, in practice, there are additional challenges when building a data space and especially for using the data to get the best value out of it. Previous efforts, e.g. FIESTA-IoT and SynchroniCity, have shown that value can be created, e.g. for experimentation [17] and analytics as a service [7], but further steps are still needed. We ascribe these to two open challenges:

- *Data Interoperability.* Although the basis for achieving data interoperability is already in place in the form of APIs and data models, the integration effort for legacy systems is still very high.
- *Data Value.* Even if the data is integrated and accessible through a single API, typically additional analytics are needed to generate valuable insights, e.g., for training a machine learning model, a data scientist still needs to work on aspects such as feature selection, feature engineering, and data cleaning.

We will discuss these two open challenges in the next two subsections, proposing possible solutions based on cutting-edge research.

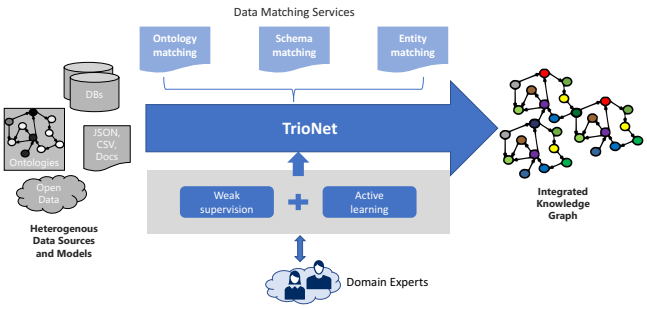

Figure 3: TrioNet: An interactive data harmonization system.

4.1 Challenge: Data interoperability

In the past, many systems or platforms were built up individually, for different application purposes and, very often, each of them behaved as an isolated data silo. This happens across different domains and organizations, and even across different departments and teams within the same organization. In a data space, there is a strong need to achieve data interoperability between these silos through an alignment of the underlying data and data models.

Towards full data interoperability, different matching/alignment problems need to be solved depending on the underlying data [23]. Fig. 3 depicts our approach to address this challenge. *TrioNet* is an interactive data integration system that utilizes weak-supervision [19] together with active learning in order to facilitate several data matching steps in a semi-automated fashion with minimal human input. TrioNet supports the data integration steps of *ontology matching, schema matching* and *entity matching*.

In ontology matching, the goal is to find semantic mappings between the elements from multiple ontologies. A widely used way to represent ontologies is standardized in the Web Ontology Language (OWL). OWL represents machine-readable knowledge about concepts and their relations to be defined in the unified form of subject-predicate-object triples. On top of that, OWL adds semantics to the underlying concepts with more general logic relations

(e.g., equivalence or disjoint relations) and constraints. In the OWL data model, each ontology O defines a, usually hierarchical, conceptualization of a domain with classes (representing domain concepts), properties (defining relations), individuals (instances of classes), and data/literal values (e.g., age of a person). Properties can define relations between classes/instances (object properties), relations to a literal (data property, defining an attribute) or relations to metadata (annotation properties) [10].

Not all data is represented in an OWL ontology. In schema matching (or schema annotation), we therefore aim to find matches between a data schema (e.g., derived from a relational database, or from CSV or JSON data) and a global, "backbone ontology" that is used to model data in context of the data space. This global ontology can, for example, be based on an agreed standard (e.g., the Smart Data Models) or constructed through a previously performed ontology matching (interlinking) between related ontologies. TrioNet can match schema to link data to the backbone ontology and then automatically transform it to the ontology format.

Last, data originating in different organizations might sometimes refer to the same real-world entity, but use a different entity representation. For instance, two organizations may use different languages or have different spelling conventions for the same physical entity. In entity matching, we aim to solve these problems by automating the task of finding such matching entities. Subsequently, the user can take a decision in terms of how to merge or link two matching entities.

4.1.1 Technical solution for Green Twin data interoperability. To achieve data interoperability in the context of Green Twin, it is necessary to provide means of data harmonization for the various sources of data, as shown in Fig. 4. However, the different data ownership and varying data privacy requirements by the involved organizations constitute a big obstacle towards achieving data harmonization. Organizations and companies do not want to share data without any control mechanisms, due to the competitive advantage associated with it. Thus, practically, data integration needs to occur in two steps: (1) Integration on a semantic, ontology/schema level, without sharing confidential information about the underlying data records and (2) full data integration, including entity matching, based on the identified alignments in (1) and after resolving data sovereignty issues (e.g., after the negotiation described in Section 3.1). Thus, our envisioned solution to this problem is to integrate data across organizations first at the schema level, using ontologies, specified in OWL to avoid the immediate sharing of confidential instance data.

4.2 Challenge: Data value

Data spaces, through capabilities for secure and trustworthy data sharing as well as interoperability, provide the technical foundation for the data economy. An additional important building block on top of the data exchange and interoperability layers will be advanced data processing functions that have a high level of re-usability and require minimum manual configuration effort. In a mature data ecosystem, multiple vendors will offer commercial functions for data processing, including but not limited to prediction, simulation, and optimization. We restrict the discussion in this article to machine learning-based prediction functions, but similar observations hold for other classes of processing functions as well.

The landscape of today's readily available functions can be separated into (a) vertical solutions for specific tasks in specific domains and (b) general-purpose machine learning libraries. Vertical solutions come in various flavors, including trained models ready for application, pre-trained models to be fine-tuned with additional data, and untrained models whose architecture is specialized to the prediction target and available input data. General purpose libraries such as sklearn or TensorFlow can be used to train models for a wide range of tasks; given a dataset from a specific application, data scientist have a wide range of possible preprocessing functions and trainable models available to build accurate data-driven prediction models. Nowadays, sophisticated methods to even automate the process of model selection and model configuration (e.g., AutoML) are available, which reduces the human effort to identification of the most suitable model input and training data, as well as to final model verification and deployment.

By leveraging the data space layer of semantic understanding and data interoperability as described in the preceding section, the gap between application-specific functions (with a single purpose) and generic functions (requiring effort and expertise to specialize) can be substantially narrowed. Having explicit knowledge of what the different pieces of data represent, and how they are semantically connected, helps in various ways. One direction is to start from application-specific models and broaden their reach to equivalent applications, automatically transferring the knowledge about required input data sources to the new task. More specifically, developers can specify the input data of a model (both for training and for prediction) in terms of its *application context*. Mathematically speaking, the application context can be represented as a node in a knowledge graph, and the set of input data sources are determined by a specific *graph neighborhood*. One realization of that idea has been provided by the CASTOR platform [12] for the Internet-of-Things, and validated by provisioning of replicable prediction models for power networks.

The power of semantic annotations of data sources can be further exploited by modularization of the process of model creation. In the absence of ground-truth labels for training a model, the principle of *data programming* [19] can help to provide an ensemble of noisy labels. A catalogue of labeling functions can be made available as a service, and the quality of each labeling function can be automatically assessed by transfer of knowledge from semantically similar classification tasks where ground truth is available.

Starting from the side of general-purpose libraries with automatic model selection and configuration (AutoML), the semantic interoperability layer opens the potential for full automation of even the data collection process. Given an explicit graph of data sources, their semantics, and their mutual relationships, the graph neighborhood of a data source provides an excellent set of candidate inputs for a machine learning model that predicts related properties. The capability to build prediction models fully automatically on top of the existing data space will be a key enabler to make predictions a basic commodity that can be set up as easily as making a conventional data subscription. This applies to the following categories of predictions:

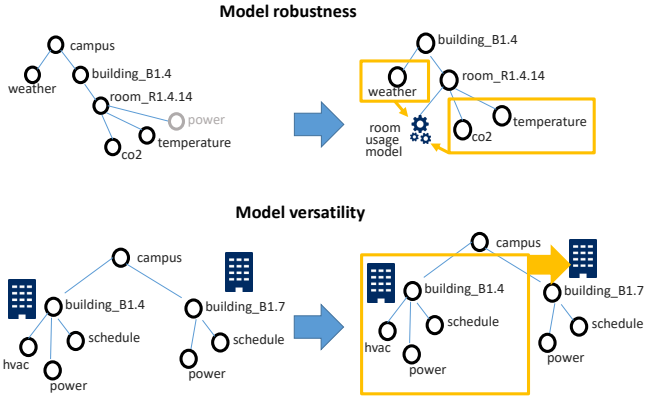

Figure 4: Generating data value through robust and versatile machine learning models.

- Predicting future values of attributes. Such predictions can be commissioned with a single command that has two prediction-related parameters: (a) the attribute to predict, and (b) the future time horizon to predict.
- Predicting the current value of an attribute from other data sources. Such predictions are useful for dealing with unreliable data sources (e.g. sensors) which can stop providing data at any time.

To summarize the considerations so far, the semantic data interoperability layer of data spaces enables a novel type class of machine learning functions that are characterized by (a) high reusability across similar tasks, and (b) minimum manual effort for data curation and model selection. Existing concepts like AutoML, multi-task learning and data programming are key enablers, but a number of additional challenges need to be solved for the new *data value layer* to become a reality:

- *Adaptivity of models in face of dynamic data sources:* Data spaces are characterized by ever-changing availability of data sources. The key principle of data sovereignty has the technical implication that data access can be granted and revoked at any time. In addition, Internet-of-Things data sources such as sensors are likely to become faulty or simply unavailable over time, whereas new data sources might be added. This dynamicity influences not only the set of available model inputs at prediction time, but also the ability to (re-) train the models. Thus, models need to be created with a great flexibility regarding their inputs, degrading gracefully as data sources become unavailable.
- *Transfer of knowledge across the semantic graph:* A model that predicts properties of a specific semantic entity, using information from the graph neighborhood as model input, can be transferred to other entities of the same type. The challenge is that there is no guarantee that the new entity has an equivalent neighborhood of data sources. Some data sources might be missing (e.g., due to sensors not installed), or additional data sources might be available as valuable inputs. The ability for transferring knowledge across models in light of heterogeneous neighborhoods is an unsolved challenge.
- *Explicit handling of limited modeling quality:* A consequence of the volatility of data sources is that the quality of predictions is never guaranteed. Even with model quality degrading in a graceful manner, the downstream applications need to be aware of new inaccuracies resulting, e.g., for one or more input data sources becoming unavailable. Thus, models at any time need to provide information not only about predicted values, but also about the certainty of prediction. However, estimating the certainty is a non-trivial task in the situation of changing availability of data sources, which comes on top of the common issues such as data drift and over- or under-fitting.

4.2.1 Technical solution for Green Twin data value. As described in section 2 the represented reality of the Green Twin use case is of a large complexity. The knowledge graph of the Green Twin is composed by multiple entities of the same type, such as multiple buildings, multiple rooms, multiple streets.

The feature set of a single entity, such as a building, varies with time (e.g., sensor faults, network faults, mobile sensors). Thus, the implementation of the analytics function is not trivial. The exploitation of automatic feature selection might help in these cases to have a reliable set of feature for a long-enough time interval. A complementary solution is the exploitation of the semantic graph to reconstruct or substitute the missing features. For example, the missing information of a room occupancy for a specific room might be substituted or reconstructed from the room occupancy of an adjacent room with the same purpose (e.g., lecture room) in the same building.

Following this approach, we might exploit a similar solution for the transfer of knowledge of an analytics model from an entity to another entity. For example, an energy consumption prediction model of a building might be transferred to another building. Two buildings might have different feature sets with different data quality (e.g., granularity). A full time series can be generated through the prediction of values within a certain time window, and this data can be used as input for the transferred model.

5 CONCLUSION

This paper describes how to enable the data spaces for the future data economy. The groundwork for realizing data spaces is currently ongoing through various research, development, and standard activities. This paper highlights the existing assets and technical developments for data spaces in IDSA, Gaia-X, and FIWARE, as well as the two key challenges, namely data interoperability and data value. These open challenges should be addressed to enable data spaces which can generate value to the stakeholders in various business domains of the data economy.

Acknowledgments: This work has been partially supported by the project SALTED (Situation-Aware Linked heterogeneous Enriched Data) from the European Union's Connecting Europe Facility programme under the Action Number 2020-EU-IA-0274, and by the Spanish State Research Agency (AEI) by means of the project SITED (Semantically-enabled Interoperable Trustworthy Enriched Data-spaces) under Grant Agreement No. PID2021-125725OB-I00.

REFERENCES

[1] International Data Spaces Association. 2021. Gaia-X and IDS. *Position Paper, Version 1.0* 01 (2021).

[2] International Data Spaces Association. 2021. IDSA Reference Architecture Model 3.0. *IDSA* 07 (2021).

[3] International Data Spaces Association. 2021. International Data Spaces Enabling Data Economy. *IDSA Brochure* 09 (2021).

[4] International Data Spaces Association. 2022. International Data Spaces Global: IDS-G. https://github.com/International-Data-Spaces-Association/IDS-G.

[5] Simona Autolitano and Agnieszka Pawlowska. 2021. Europe's quest for digital sovereignty: GAIA-X as a case study. *IAI Papers* 21, 14 (2021), 1–22.

[6] Arnaud Braud, Gaël Fromentoux, Benoit Radier, and Olivier Le Grand. 2021. The road to European digital sovereignty with Gaia-X and IDSA. *IEEE Network* 35, 2 (2021), 4–5.

[7] Flavio Cirillo, David Gómez, Luis Diez, Ignacio Elicegui Maestro, Thomas Barrie Juel Gilbert, and Reza Akhavan. 2020. Smart city IoT services creation through large-scale collaboration. *IEEE Internet of Things Journal* 7, 6 (2020), 5267–5275.

[8] Flavio Cirillo, Gürkan Solmaz, Everton Luís Berz, Martin Bauer, Bin Cheng, and Ernoe Kovacs. 2019. A standard-based open source IoT platform: FIWARE. *IEEE Internet of Things Magazine* 2, 3 (2019), 12–18.

[9] Context Information Management (CIM) ETSI Industry Specification Group (ISG). 2021. *NGSI-LD API.* Technical Report. https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.04.01_60/gs_cim009v010401p.pdf

[10] Diego De Uña, Nataliia Rümmele, Graeme Gange, Peter Schachte, and Peter J Stuckey. 2018. Machine Learning and Constraint Programming for Relational-To-Ontology Schema Mapping.. In *IJCAI*, Vol. 2018. 27th.

[11] DSBA. 2022. The Data Spaces Business Alliance - unleashing the European Data Economy. https://data-spaces-business-alliance.eu/. (Accessed on 09/23/2022).

[12] Bradley Eck, Francesco Fusco, Robert Gormally, Mark Purcell, and Seshu Tirupathi. 2020. Scalable deployment of AI time-series models for IoT. *arXiv preprint arXiv:2003.12141* (2020).

[13] Günter Eggers, Bernd Fondermann, Berthold Maier, Klaus Ottradovetz, Julius Pfrommer, Ronny Reinhardt, Hannes Rollin, Arne Schmieg, Sebastian Steinbuß, Philipp Trinius, et al. 2020. GAIA-X: Technical Architecture. *Federal Ministry for Economic Affairs and Energy (BMWi) Public Relations Division, Berlin* 06 (2020).

[14] FIWARE. 2020. Scorpio Broker. https://scorpio.readthedocs.io/en/latest/. (Accessed on 09/21/2022).

[15] FIWARE Foundation, TMForum, IUDX, and OASC. 2020. Smart Data Models – A global program. https://smartdatamodels.org/. (Accessed on 09/21/2022).

[16] Gaia-X. 2022. Gaia-X - Architecture Document - 22.04 Release. 04 (2022).

[17] Jorge Lanza, Luis Sanchez, Juan Ramon Santana, Rachit Agarwal, Nikolaos Kefalakis, Paul Grace, Tarek Elsaleh, Mengxuan Zhao, Elias Tragos, Hung Nguyen, et al. 2018. Experimentation as a service over semantically interoperable Internet of Things testbeds. *IEEE Access* 6 (2018), 51607–51625.

[18] Shiori Ota and M Sc Mikkel Knudsen. 2021. Exploring Japan's Society 5.0. (09 2021).

[19] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 269.

[20] Akira Sakaino. 2022. International Collaboration Between Data Spaces and Carrier Networks. (2022), 471–483.

[21] Luis Sánchez, Jorge Lanza, and Luis Muñoz. 2020. From the Internet of Things to the Social Innovation and the Economy of Data. *Wireless Personal Communications* 113, 3 (2020), 1407–1421.

[22] Pablo Sotres, Jorge Lanza, Luis Sánchez, Juan Ramón Santana, Carmen López, and Luis Muñoz. 2019. Breaking vendors and city locks through a semantic-enabled global interoperable internet-of-things system: A smart parking case. *Sensors* 19, 2 (2019), 229.

[23] Saravanan Thirumuruganathan, Nan Tang, Mourad Ouzzani, and AnHai Doan. 2020. Data Curation with Deep Learning.. In *EDBT*. 277–286.

[24] International Data Spaces Association version 4.1.0. 2021. International Data Spaces Information Model. https://w3id.org/idsa/core. (Accessed on 09/16/2022).

[25] W3C. 2022. Decentralized Identifiers (DIDs) v1.0, Core architecture, data model, and representations. (Accessed on 09/23/2022).

[26] W3C. 2022. Verifiable Credentials Data Model v1.1. (Accessed on 09/23/2022).