Chapter 4
Semantic Intelligence in Big Data Applications

Valentina Janev

Abstract Today, data are growing at a tremendous rate, and according to the International Data Corporation, it is expected they will reach 175 zettabytes by 2025. The International Data Corporation also forecasts that more than 150B devices will be connected across the globe by 2025, most of which will be creating data in real time, while 90 zettabytes of data will be created by Internet of things (IoT) devices. This vast amount of data creates several new opportunities for modern enterprises, especially for analyzing enterprise value chains in a broader sense. In order to leverage the potential of real data and build smart applications on top of sensory data, IoT-based systems integrate domain knowledge and context-relevant information. Semantic intelligence is the process of bridging the semantic gap between human and computer comprehension by teaching a machine to think in terms of object-oriented concepts in the same way as a human does. Semantic intelligence technologies are the most important component in developing artificially intelligent knowledge-based systems, since they assist machines in contextually and intelligently integrating and processing resources. This chapter aims at demystifying semantic intelligence in distributed, enterprise, and Web-based information systems. It also discusses prominent tools that leverage semantics, handle large data at scale, and address challenges (e.g., heterogeneity, interoperability, and machine learning explainability) in different industrial applications.

Keywords Semantic intelligence · Big data applications · Knowledge graphs · Artificial intelligence · Interoperability

Key Points

- Semantic intelligence is the process of bridging the semantic gap between human and computer comprehension.

V. Janev (✉)
Institute Mihajlo Pupin, University of Belgrade, Belgrade, Serbia
e-mail: valentina.janev@institutepupin.com

S. Jain, S. Murugesan (eds.), *Smart Connected World*,
https://doi.org/10.1007/978-3-030-76387-9_4

- There is a need for semantic standards to improve the interoperability of complex systems.
- The semantic data lakes supply the data lake with a semantic middleware that allows uniform access to original heterogeneous data sources.
- Knowledge graphs is a solution that allows the building of a common understanding of heterogeneous, distributed data in organizations and value chains, and thus provision of smart data for artificial intelligence applications.
- The goal of semantic intelligence is to make business intelligence solutions accessible and understandable to humans.

4.1 Introduction

Both researchers and information technology (IT) professionals have to cope with a large number of technologies, frameworks, tools, and standards for the development of enterprise Web-based applications. This task has become even more cumbersome as a result of the following events:

- The emergence of the Internet of things (IoT) in 1999 (Rahman & Asyhari, 2019)
- The development of Semantic Web (SW) technologies as a cornerstone for further development of the Web (Berners-Lee, 2001)
- The development of big data solutions (Laney, 2001)

Hence, topics such as smart data management (Alvarez, 2020), linked open data (Auer et al., 2014), semantic technologies (Janev & Vraneš, 2009), and smart AU1 analytics have spawned a tremendous amount of attention among scientists, software experts, industry leaders, and decision-makers. Table 4.1 defines a few terms related to data, such as open data, big data, linked data, and smart data.

Table 4.1 Definitions

Term	Definition
Open data	"The data available for reuse free of charge can be observed as open data" (Janev et al., 2018)
Big data	"'Big data' are high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Laney, 2001)
	"Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization" (Manyika, 2011)
Linked data	The term "linked data" refers to a set of best practices for publishing structured data on the Web. These principles have been coined by Tim Berners-Lee in the design issue note Linked Data[a] (Berners-Lee, 2006)
Smart data	"Simply put, if big data are a massive amount of digital information, smart data are the part of that information that is actionable and makes sense. It is a concept that developed along with, and thanks to, the development of algorithm-based technologies, such as artificial intelligence and machine learning" (Dallemand, 2020)

[a]https://www.w3.org/DesignIssues/LinkedData

Despite the fact that the term IoT ("sensors and actuators embedded in physical objects and connected via wired and wireless networks") is 20 years old, the actual idea of connected devices is older and dates back to the 1970s. In the last two decades, with the advancement in ITs, new approaches have been elaborated and tested for handling the influx of data coming from IoT devices. On one side, the focus in industry has been on manufacturing and producing the right types of hardware to support IoT solutions. On the other, the software industry is concerned with finding solutions that address issues with different aspects (dimensions) of data generated from IoT networks, including (1) the *volume* of data generated by IoT networks and the methods of storing data, (2) the *velocity* of data and the speed of processing, and (3) the *variety* of (unstructured) data that are communicated via different protocols and the need for adoption of standards. While these three Vs have been continuously used to describe big data, additional dimensions have been added to describe data integrity and quality, such as (4) *veracity* (i.e., truthfulness or uncertainty of data, authenticity, provenance, and accountability), (5) *validity* (i.e., correct processing of data), (6) *variability* (i.e., context of data), (7) *viscosity* (i.e., latency data transmission between the source and destination), (8) *virality* (i.e., speed of the data sent and received from various sources), (9) *vulnerability* (i.e., security and privacy concerns associated with data processing), (10) *visualization* (i.e., interpretation of data and identification of the most relevant information for the users), and (11) *value* (i.e., usefulness and relevance of the extracted data in making decisions and capacity to turn information into action).

With the rapid development of the IoT, different technologies have emerged to bring the knowledge (Patel et al., 2018) within IoT infrastructures to better meet the purpose of the IoT systems and support critical decision-making (Ge et al., 2018; Jain, 2021). While the term "big data" refers to datasets that have large sizes and complex structures, the term "big data analytics" refers to the strategy of analyzing large volumes of data which are gathered from a wide variety of sources, including different kind of sensors, images/videos/media, social networks, and transaction records. Aside from the analytic aspect, big data technologies include numerous components, methods, and techniques, each employed for a slightly different purpose, for instance for pre-processing, data cleaning and transformation, data storage, and visualization.

In addition to the emergence of big data, the last decade has also witnessed a technology boost for artificial intelligence (AI)-driven technologies. A key prerequisite for realizing the next wave of AI application is to leverage data, which are heterogeneous and distributed among multiple hosts at different locations. Consequently, the fusion of big data and IoT technologies and recent advancements in machine learning have brought renewed visibility to AI and have created opportunities for the development of services for many complex systems in different industries (Mijović et al., 2019; Tiwari et al., 2018). Nowadays, it is generally accepted that AI methods and technologies bring transformative change to societies and industries worldwide. In order to reduce the latency, smart sensors (sensor networks) are empowered with embedded intelligence that performs pre-processing, reduces the volume, and reacts autonomously. Additionally, in order to put the data in context, standard data models are associated with data processing services, thus facilitating the deployment of sensors and services in different environments.

This chapter explains the need for semantic standards that improve interoperability in complex systems, introduce the semantic lake concept, and demystify the semantic intelligence in distributed, enterprise, and Web-based information systems (see the following section). In order to select an appropriate semantic description, processing model, and architecture solution, data architects and engineers need to become familiar with the analytical problem and the business objectives of the targeted application. Therefore, the authors describe four eras of data analytics and introduce different big data tools.

4.2 From Data to Big Data to Smart Data Processing

Data-driven technologies such as big data and the IoT, in combination with smart infrastructures for management and analytics, are rapidly creating significant opportunities for enhancing industrial productivity and citizen quality of life. As data become increasingly available (e.g., from social media, weblogs, and IoT sensors), the challenge of managing them (i.e., selecting, combining, storing, and analyzing them) is growing more urgent (Janev, 2020). Thus, there is a demand for development of computational methods for the ingestion, management, and analysis of big data, as well as for the transformation of these data into knowledge.

From a data analytics point of view, this means that data processing has to be designed taking into consideration the diversity and scalability requirements of the targeted domain. Furthermore, in modern settings, data acquisition occurs in near real time (e.g., IoT data streams), and the collected and pre-processed data are combined with batch loads by different automated processes. Hence, novel architectures are needed; these architectures have to be "flexible enough to support different service levels as well as optimal algorithms and techniques for the different query workloads" (Thusoo et al., 2010).

4.2.1 Variety of Data Sources

The development of big data-driven pipelines for transforming big data into actionable knowledge requires the design and implementation of adequate IoT and big data processing architecture, where, in addition to volume and velocity, the variety of available data sources should be considered. The processing and storage of data which are generated by a variety of sources (e.g., sensors, smart devices, and social media in raw, semi-structured, unstructured, and rich media formats) is complicated. Hence, different solutions for distributed storage, cloud computing, and data fusion are needed (Liu et al., 2015). In order to make the data useful for data analysis, companies use different methods to reduce complexity, downsize the data scale (e.g., dimensional reduction, sampling, and coding), and pre-process the data (i.e., data extraction, data cleaning, data integration, and data transformation) (Wang, 2017). Data heterogeneity can thus be defined in terms of several dimensions:

- *Structural variety*, which refers to data representation and indicates multiple data formats and models. For instance, the format of satellite images is very different from the format used to store tweets which are generated on the Web.
- *Media variety*, which refers to the medium in which data get delivered. For instance, the audio of a speech vs. the transcript of the speech may represent the same information in two different media.
- *Semantic variety*, which refers to the meaning of the units (terms) used to measure or describe the data that are needed to interpret or operate on the data. For instance, a standard unit for measuring electricity is the kilowatt; however, the electricity generation capacity of big power plants is measured in multiples of kilowatts, such as megawatts and gigawatts.
- *Availability variations*, which mean that the data can be accessed continuously (e.g., from traffic cameras) or intermediately (e.g., only when the satellite is over the region of interest).

In order to enable broad data integration, data exchange, and interoperability, and to ensure extraction of information and knowledge, standardization at different levels (e.g., metadata schemata, data representation formats, and licensing conditions of open data) is needed. This encompasses all forms of (multilingual) data, including structured and unstructured data, as well as data from a wide range of domains, including geospatial data, statistical data, weather data, public sector information, and research data, to name a few.

4.2.2 The Need for Semantic Standards

In 1883, Michel Bréal, a French philologist, coined the term "semantics" to explain how terms may have various meanings for different people, depending on their experiences and emotions. In the information processing context, semantics refers to the "meaning and practical use of data" (Woods, 1975), namely, the efficient use of a data object for representing a concept or object. Since 1980, the AI community has promoted the concept of providing general, formalized knowledge of the world to intelligent systems and agents (see also the panel report from the 1997 *Data Semantics: what, where and how?*) (Sheth, 1997).

In 2001, Sir Tim Berners-Lee, Director of the World Wide Web Consortium (W3C), presented his vision for the SW, describing it as an expansion of the traditional Web and a global distributed architecture where data and services can easily interact. In 2006, Berners-Lee also introduced the basic (linked data) principles for interlinking datasets on the Web via references to common concepts. The Resource Description Framework (RDF) norm is used to reflect the knowledge that defines the concepts. Parallel to this, increased functionalities and improved t2.1 **Table 4.2** An overview of (recommended) Semantic Web technologies

	Technology	Definition
t2.2	Technology	Definition
t2.3	RDF, 2004	RDF is a general-purpose language for encoding and representing data on the Internet The RDF Schema is used to represent knowledge in terms of objects ("resources") and relationships between them
t2.4	RDFS, 2004	RDF Schema serves as the meta language or vocabulary to define properties and classes of RDF resources
t2.5	SPARQL, 2008	SPARQL Query Language for RDF is a standard language for querying RDF data
t2.6	OWL, 2004	OWL is a standard Web Ontology Language that facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF-S by providing additional vocabulary along with a formal semantics
t2.7	SWRL, 2004	SWRL aims to be the standard rule language of the Semantic Web. It is based on a combination of the OWL DL, OWL Lite, RuleML, etc.
t2.8	WSDL, 2007	WSDL provides a model and an XML format for describing Web services
t2.9	SAWSDL, 2007	SAWSDL (Semantic Annotations for WSDL and XML Schema) explains how to apply semantic annotations to WSDL and XML Schema documents
t2.10	RDFa, 2008	A collection of attributes and processing rules for extending XHTML to support RDF
t2.11	GRRDL, 2007	A mechanism for Gleaning Resource Descriptions from Dialects of Languages (e.g., microformats)
t2.12	OWL 2, 2012	OWL 2 extends the W3C OWL Web Ontology Language with a small but useful set of features (EL, QL, RL) that enable effective reasoning
t2.13	DQV, 2015	Data Quality Vocabulary is an extension to the DCAT vocabulary to cover the quality of the data
t2.14	SHACL, 2017	Shapes Constraint Language is a language for validating RDF graphs against a set of conditions
t2.15	DCAT, 2020	Data Catalog Vocabulary is an RDF vocabulary for facilitating interoperability between Web-based data catalogs

robustness of modern RDF stores, as well as wider adoption of standards for representing and querying semantic knowledge, such as RDF(s) and SPARQL, have adopted linked data principles and semantic technologies in data and knowledge management tasks. Table 4.2 gives an overview of (recommended) SW technologies by the W3C.[1]

Aside from the W3C, there are a few international organizations (associations or consortia) that are important for assessing and standardizing ITs, such as IEEE-SA (see The Institute of Electrical and Electronics Engineers Standards Association[2]), OASIS (see The Organization for the Advancement of Structured Information Standards[3]), and a number of others.

[1] http://www.w3.org/.
[2] http://standards.ieee.org/.
[3] http://www.oasis-open.org/.

4.2.3 Semantic Integration and Semantic Data Lake Concept

In Tim Berners's vision, the Web is a massive platform-neutral engineering solution that is service-oriented, with service specified by machine-processable metadata, formally defined in terms of messages which are exchanged between provider and requester agents, rather than the properties of the agents themselves. In the last 10 years, businesses have embraced Tim Berners's vision and the linked data approach, and cloud computing infrastructures have enabled the emergence of semantic data lakes.

The following are some of the ways by which computer scientists and software providers have tackled the emerging problems in the design of end-to-end data/knowledge processing pipelines:

- In addition to operational database management systems (present on the market since the 1970s), different NoSQL stores appeared that lack adherence to the time-honored SQL principles of ACID (i.e., atomicity, consistency, isolation, and durability) (Table 4.3).
- Cloud computing emerged as a paradigm that focuses on sharing data and computations over a scalable network of nodes including end user computers, data centers, and Web services (Assunção et al., 2015).
- The concept of open data emerged ("data or content that anyone is free to use, reuse and redistribute") as an initiative to enable businesses to use open data sources to improve their business models and drive a competitive advantage (see an example of integrating open data in end-to-end processing in modern ecosystem in Fig. 4.1).
- The concept of data lake as a new storage architecture was promoted; in it, raw data can be stored regardless of source, structure, and (usually) size. As a result, the data warehousing method (which is built on a repository of centralized, filtered data that have already been processed for a particular purpose) is seen as obsolete, as it causes problems with data integration and adding new data sources.

The development of business intelligence services is simple, when all data sources collect information based on unified file formats and the data are uploaded to a data warehouse. However, the biggest challenge that enterprises face is the undefined and unpredictable nature of data appearing in multiple formats. Additionally, in order to gain competitive advantage over their business rivals, the companies utilize open data resources that are free from restrictions, can be reused and redistributed, and can provide immediate information and insights. Thus, in a modern data ecosystem, data lakes and data warehouses are both widely used for storing big data. A data warehouse (Kern et al., 2020) is a repository for structured, filtered data that have already been processed for a specific purpose. A data lake is a large, raw data repository that stores and manages the company's data bearing any format. Moreover, recently, *semantic data lakes* (Mami et al., 2019) were introduced as an extension of the data lake supplying it with a semantic middleware, which AU3 t3.1 **Table 4.3** Semantic intelligence in the drug domain (example)

		Step	Description
t3.2		Step	Description
t3.3	1	Identification of datasets	The data architect first identifies the existing company data sources, as well as available open data sources (e.g., DrugBank and DBpedia)
t3.4		Elaboration of business questions	The business users specify questions to be answered with a unified access interface to a set of autonomous, distributed, and heterogeneous data sources, as well as with AI-based business intelligence services
t3.5	2	Development of semantic models	In the case of the drug domain, the drug dataset has properties such as generic drug name, code, active substances, non-proprietary name, strength value, cost per unit, manufacturer, related drug, description, URL, and license. Hence, ontology development can leverage reuse of classes and properties from existing ontologies and vocabularies including Schema.org vocabulary[a], DBpedia Ontology[b], UMBEL (Upper Mapping and Binding Exchange Layer)[c], DICOM (Digital Imaging and Communications in Medicine)[d], and DrugBank
t3.6	3	Elaboration of extraction rules	The data administrator runs the extraction process using software tools, such as OpenRefine (which the authors used), RDF Mapping Language[e], and XLWrap[f], which is a Spreadsheet-to-RDF Wrapper, among others
t3.7		Elaboration of mapping rules	For the identified datasets (i.e., Excel, XLS data, and MySQL store), the data administrator can specify and run mapping rules in order to query the data on-the-fly without data transformation or materialization
t3.8	4	Elaboration of quality assessment services	The business user/data architect specifies models for describing the quality of the semantic (big linked) data which are needed. Zaveri et al. (2016), for instance, grouped the dimensions into: • *Accessibility*: availability, licensing, interlinking, security, and performance • *Intrinsic*: syntactic validity, semantic accuracy, consistency, conciseness, and completeness • *Contextual*: relevancy, trustworthiness, understandability, and timeliness • *Representational*: representational conciseness, interoperability, interpretability, and versatility
t3.9	5	Standardization of interlinking	Specialized tools are used to help the interlinking and to discover links between the source and target datasets. Since the manual mode is tedious, error-prone, and time-consuming, and the fully automated mode is currently unavailable, the semi-automated mode is preferred and reliable. Link generation application yields links in RDF format using *rdfs:seeAlso* or *owl:sameAs* predicates
		Standardization of data querying connectors	The data administrator specifies connectors as standardized components for interoperability between different solutions. Once the datasets are prepared based on standard vocabularies, the next step is to provide standard querying mechanisms. To this aim, vocabularies such as DCAT and DQV are used to

(continued)

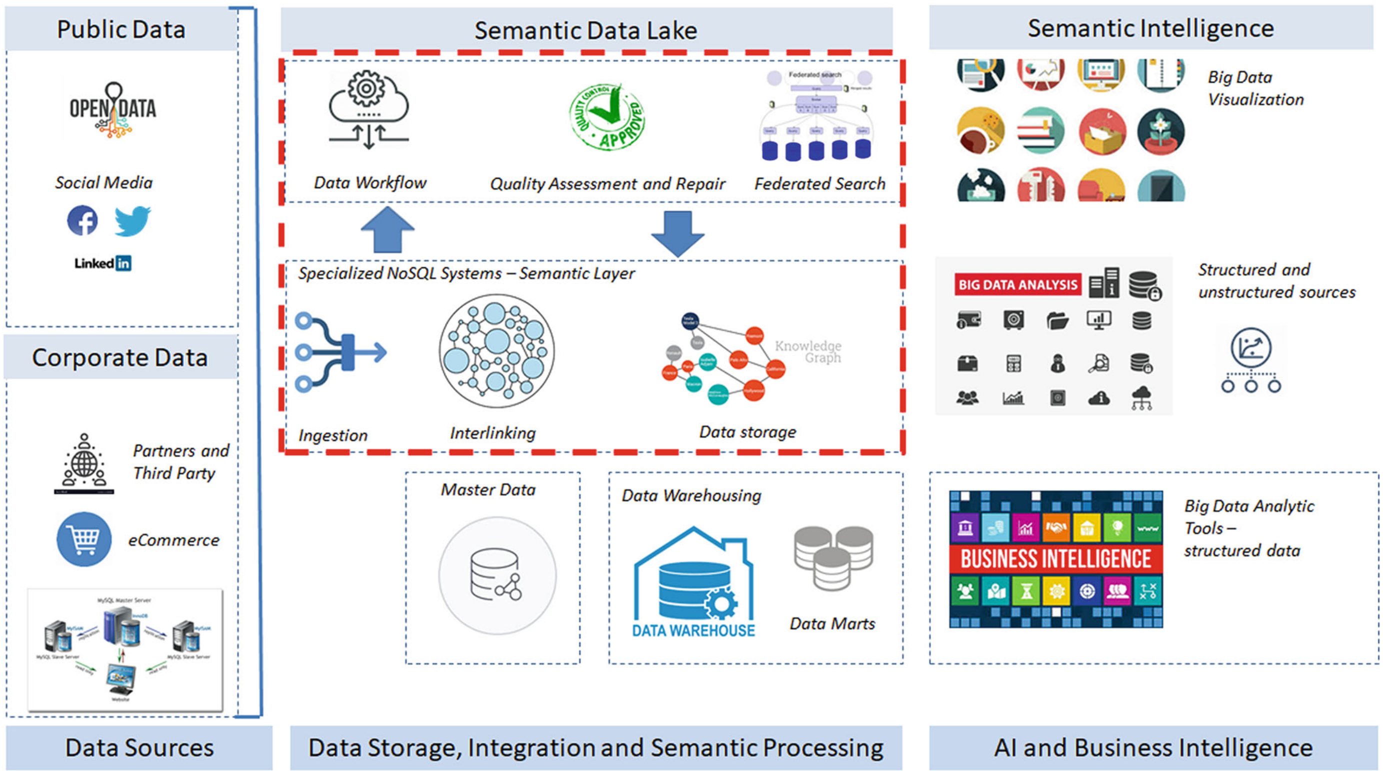
Table 4.3 (continued)

	Step	Description
		describe the datasets and standardize the access to data. SPARQL is one of the standard querying languages for RDF KGs
6	Exploration via federated querying	Intelligently searching vast datasets of drug data (i.e., patents, scientific publications, and clinical trials) data will help, for instance, accelerate the discovery of new drugs and gain insights into which avenues are likely to yield the best results. Federated query processing techniques (Endris et al., 2020) provide a solution to scale up to large volumes of data distributed across multiple data sources. Source details are used to find efficient execution plans that reduce the overall execution time of a query while increasing the completeness of the answers
7	Advanced Data Analytics Services	Drug data aggregated with other biomedical data often display different levels of granularity, that is, a variety of data dimensionalities, sample sizes, sources, and formats. In order to support human decision-making, different widgets are needed for visualization and tracing the results of interactive analysis
	Advanced Business Intelligence Services	Algorithm-based techniques (i.e., machine learning and deep learning algorithms) have already been used in drug discovery, bioinformatics, and cheminformatics. What is new in semantic intelligence-based systems is that contextual information from the KG can be used in machine learning, thus improving, for instance, the recommendation and explainability capabilities (Fletcher, 2019; Patel et al., 2020)
8	Integration in big data ecosystem	There are multiple ways of exposing and exploring the KGs-based services to public and other businesses, for instance, using the *data-as-a-service* or *software-as-a-service* concept

[a] https://schema.org/
[b] https://wiki.dbpedia.org/services-resources/ontology
[c] http://umbel.org/
[d] https://www.dicomstandard.org/
[e] https://github.com/RMLio
[f] http://xlwrap.sourceforge.net/

allows uniform access to original heterogeneous data sources. *Semantic data lakes* integrate knowledge graphs (KGs), a solution that allows the building of a common understanding of heterogeneous, distributed data in organizations and value chains, and thus provision of smart data for AI applications.

In 2012, the announcement of the Google Knowledge Graph drew much attention to graph representations of general world knowledge. In the last decade, enterprise settings have shown a tendency to collect and encapsulate metadata in a form of corporate knowledge (or smart data) using semantic technologies, while the data are stored or managed via an enterprise KG. However, many factors have prevented effective large-scale development and implementation of complex knowledge-based



AU2

Fig. 4.1 Modern data ecosystem

scenarios because of the inability to cope with the rising challenges coming from big data applications, the rigidity of existing database management systems, the inability to go beyond the standard requirements of query answering, and the lack of knowledge languages expressive enough to address real-world cases. Despite the challenges, the voluntary created KGs such as DBpedia (Auer et al., 2007a, b) motivated many big companies (e.g., Google, Facebook, and Amazon) to explore the benefits of using semantic technologies for profit.

4.3 Semantics and Data Analytics

Data analytics is a concept that refers to a group of technologies that are focused on data mining and statistical analysis. Data analytics has grown in popularity as a field of study for both practitioners and academics over the last 70 years. The Analytics 1.0 era started in the 1950s and lasted roughly 50 years. With the advent of relational databases in the 1970s and the invention of the Web by Sir Tim Berners-Lee in 1989, the data analytics progressed dramatically as a new software approach, and AI was developed as a separate scientific discipline.

The Analytics 2.0 era began in the 2000s with the introduction of Web 2.0-based social and crowdsourcing systems. Although business solutions in the Analytics 1.0 era were focused on relational and multidimensional database models, the Analytics 2.0 era introduced NoSQL and big data database models, which opened up new goals and technological possibilities for analyzing large volumes of semi-structured data. Before big data and after big data are terms companies and data scientists use to describe these two spans of time (Davenport, 2013).

The fusion of internal data with externally sourced data from the Internet, different types of sensors, public data projects (e.g., the human genome project), and captures of audio and video recordings were made possible by a new generation of tools with fast-processing engines and NoSQL stores. The data science area (a multifocal field consisting of an intersection of mathematics and statistics, computer science, and domain specific knowledge) also advanced significantly during this period, delivering scientific methods, exploratory processes, algorithms, and resources that can be used to derive knowledge and insights from data in various forms. The IoT and cloud computing technologies ushered in the Analytics 3.0 era, allowing for the creation of hybrid technology environments for data storage, real-time analysis, and intelligent customer-oriented services. After the countless possibilities for capitalizing on analytics resources, Analytics 3.0 is also known as *the era of impact* or *the era of data-enriched offerings* after the endless opportunities for capitalizing on analytics services. For creating value in the data economy, Davenport (2013) suggested that the following factors need to be properly addressed:

- Combining multiple kinds of information
- Adoption of novel information management tools

- Introduction of "agile" analytical methods and machine-learning techniques to generate insights at a much faster rate
- Embedding analytical and machine learning models into operational and decision processes
- Development of skills and processes for data exploration and discovery
- Requisite skills and processes to develop prescriptive models that involve large-scale testing and optimization and are a means of embedding analytics into key processes
- Leveraging new approaches to decision-making and management

The aim of the Analytics 4.0 era, also known as *the era of consumer-controlled data*, is to give consumers complete or partial control over data. There are various possibilities for automating and augmenting human/computer communications by integrating machine translation, smart reply, chat-bots, and virtual assistants, all of which are associated with the Industry 4.0 trend.

The selection of an appropriate semantic processing model (i.e., vocabularies, taxonomies, and ontologies that facilitate interoperability) (Mishra & Jain, 2020) and analytical solution is a challenging problem and depends on the business issues of the targeted domain, for instance, e-commerce, market intelligence, e-government, healthcare, energy efficiency, emergency management, production management, and/or security.

4.4 Semantics and Business Intelligence Applications

The topic semantic intelligence brings together the efforts of AI, machine learning, and SW communities. The choice of an effective processing model and analytical approach is a difficult task that is influenced by the business concerns of the targeted domain, for instance, risk assessment in banks and the financial sector, predictive maintenance of wind farms, sensing and cognition in production plants, and automated response in control rooms. The integration of advanced analytical services with semantic data lakes is a complex and hot research topic (see the eight-step process in Fig. 4.2). Although the aim of semantics is to make data and processes understandable to machines, the goal of semantic intelligence is to make business intelligence solutions accessible and understandable to humans. Natural language processing and semantic analysis, for example, are used to understand and address posted questions while incorporating semantic knowledge in human-machine interfaces (digital assistants). In this case, natural language processing methods combine statistical and linguistic methods with graph-based AI.

Example This example presents the process of creating and publishing a linked drug dataset based on open drug datasets from selected Arabic countries. The drug dataset has been integrated in a form of a materialized KG (Lakshen et al., 2020). The overall goal is to allow the business user to retrieve relevant information about

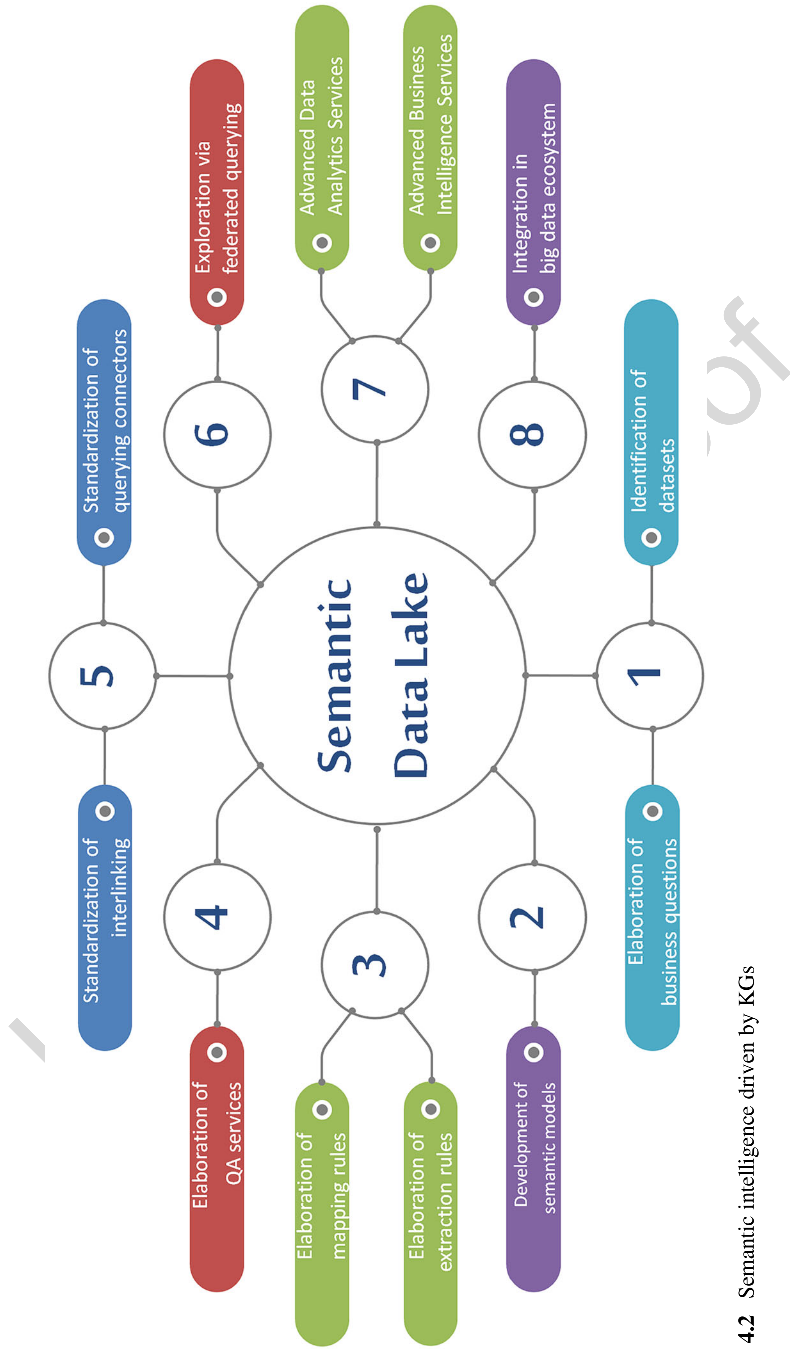


Fig. 4.2 Semantic intelligence driven by KGs

drugs from the local company data store and other open-source datasets. To this aim, an intelligent digital assistant is needed.

The pharmaceutical/drug industry was among the first that validated linked data principles and standards recommended by the W3C consortium and used the approach for precise medicine. Table 4.3 briefly describes the necessary tasks for development of a semantic data lake and leveraging AI with KGs.

4.5 Role of Semantics in (Big) Data Tools

Different keywords are used to name semantic techniques and technologies in the literature and in practice: semantic annotation tools and content indexing and categorization tools; semantic data processing and integration platforms; RDF triple storage systems; SW services (Patel & Jain, 2019) and SOA middleware platforms; semantic annotation tools, content indexing, and categorization tools; semantic search and information retrieval technologies; semantic textual similarity methods, linguistic analysis and text mining algorithms, and ontology-mediated portals; ontological querying/inference engines and rule-based engines; ontology learning methods; and ontology reasoners. In their study of the market value of semantic technologies, Davis et al. (2004) defined the following four major functions 50 commercial companies offered in 2004:

- Discover, acquire, and create semantic metadata
- Represent, organize, integrate, and inter-operate meanings and resources
- Reason, interpret, infer, and answer using semantics
- Provision, present, communicate, and act using semantics

Based on the analysis of the functionalities of more than 50 SW tools, Janev and Vraneš (2011) classified main semantic technology segments into semantic modeling and creation, semantic annotation, semantic data management and integration, semantic search and retrieval, semantic collaboration including portal technologies, and learning and reasoning. Furthermore, Janev et al. (2020) discussed challenges related to big data tools and points to a repository of big data tools; see the results of the project LAMBDA—Learning, Applying, Multiplying Big Data (Janev, 2020). We have categorized the tools into 12 categories (see also Table 4.4): Cloud Marketplaces, Hadoop as a Web Service/Platform, Operational Database Management Systems, NoSQL/Graph databases, Analytics Software/System/Platform, Data Analytics Languages, Optimization Library for Big Data, Library/API for Big Data, ML Library/API for Big Data, Visualization Software/System, and Distributed Messaging System.

The authors' analysis highlights that it is important to distinguish between big data processing, where the size (volume) is one of many important aspects of the data, and big data analytics, where semantic processing and use of semantic standards can improve the analysis and produce explainable results.

Table 4.4 Big data tools[a]

Category	Tools
Cloud marketplaces	Alibaba Cloud; IBM Cloud; Google Cloud Platform; Oracle Cloud Marketplace; CISCO Marketplace; Microsoft Azure Marketplace; AWS Marketplace
Hadoop as a Web service/ platform	HDInsight; IBM InfoSphere BigInsights; MapR; Cloudera CDH; Amazon EMR
Operational database management systems	IBM (DB2); SAP (SAP HANA); Microsoft (SQL Server); ORACLE (Database)
NoSQL/graph databases	Hadoop Distributed File System (hdfs); Amazon Neptune neo4j; TigerGraph; Mapr database; OntoText GraphDB; AllegroGraph; Virtuoso; Apache Jena; MarkLogic JanusGraph; OrientDB; Microsoft Azure Cosmos DB; Apache Hbase; Apache Cassandra; MongoDB
Stream processing engines	Apache Flume; Apache Apex; Amazon Kinesis Streams; Apache Flink; Apache Samza; Apache Storm; Apache Spark
Analytics software/system/ platform	SAS Analytics Software & solutions; MatLab; H2O.ai; Accord framework; Apache Hadoop; Cloudera data platform; VADALog system; Semantic Analytics Stack (SANSA)
Data analytics languages	Scala; Julia; SPARQL; SQL; R; Python package index (PyPI); Python
Optimization library for big data	Facebook ax; Hyperopt; IBM ILOG CPLEX optimization library
Library/API for big data	TensorFlow serving; MLLIB; BigML; Google Prediction API; Azure machine learning; Amazon machine learning API; IBM Watson programming with Big Data in R
ML library/API for big data	Caffe.ai; Apache MXNet; Xgboost; PyTorch; Keras; TensorFlow
Visualization software/ system	Oracle Visual Analyzer; Microsoft Power BI; DataWrapper; QlikView; Canvas.js; HighCharts; Fusion Chart; D3; Tableau; Google chart
Distributed messaging system	Apache Kafka

[a]LAMBDA Catalogue available at https://project-lambda.org/tools-for-experimentation

4.6 Summary

Advances in hardware and software technology, such as the IoT, mobile technologies, data storage and cloud computing, and parallel machine learning algorithms, have allowed the collection, analysis, and storage of large volumes of data from a variety of quantitative and qualitative domain-specific data sources over the last two decades. As the authors presented in this chapter, interoperable data infrastructure and standardization of data-related technology, including the creation of metadata standards for big data management, are needed to simplify and make big data processing more efficient. Semantics play an important role, particularly when it comes to harnessing domain information in the form of KGs. As the authors' analysis showed, in the last decade, especially after the announcement of the Google Knowledge Graph, large corporations introduced semantic processing technologies to provide scalable and flexible data discovery, analysis, and reporting. The semantic data lake approach has been exploited to allow uniform access to original heterogeneous data, while the semantic standards and principles are used for:

- Representing (schema and schema-less) data
- Representing metadata (about documentation, provenance, trust, accuracy, and other quality properties)
- Modeling data processes and flows (i.e., representing the entire pipeline making data representation shareable and verifiable)
- Implementing standard querying and analysis services

However, transforming big data into actionable big knowledge demands scalable methods for creating, curating, querying, and analyzing big knowledge. The authors' study on big data tools reveals that there are still open issues that impede a prevalence usage of graph-based frameworks over more traditional technologies such as relational databases and NoSQL stores. For instance, tools are needed for federations of data sources represented using the RDF graph data model for ensuring efficient and effective query processing while enforcing data access and privacy policies. Next, the integration of analytic algorithms over a federation of data sources should be assessed and evaluated. Finally, quality issues that are more likely to be present, such as inconsistency and incompleteness, should be properly addressed and integrated in the reasoning processes.

Along with the discussion of the emerging big data tools on the market (categorized into 12 groups), in this chapter, the authors summarized an eight-step approach for the utilization of KGs for semantic intelligence. Hence, it is possible to conclude that there is a broad spectrum of applications in different industries where semantic technologies and machine-learning methods are used for managing actionable knowledge in real-world scenarios.

Once the abovementioned issues are effectively addressed, promising results from semantic intelligence services and applications are expected, for instance, for personalized healthcare, financial portfolio optimization and risk management, and big data-driven energy services.

Review Questions

- What is the difference between open data, big data, linked data, and smart data?
- What are the biggest challenges that enterprises face nowadays?
- What are key requirements for development of big data-driven pipelines for transforming big data into actionable knowledge?
- How does the data analytics field develop over time?
- What is the process of development of a semantic data lake?

Discussion Questions

- How can we categorize big data tools? Which technologies are needed for transforming big data into actionable big knowledge?
- Elaborate challenges for big data ecosystems, e.g., energy domain.

- How stable are W3C standards? How often are they used for building semantic intelligence applications? Do you know other standards for building semantic applications?
- Discuss extraction rules and standards for different data sources.

Problem Statements for Young Researchers

- Compare the data warehousing and data lakes concepts.
- Discover different ways for building semantic data lakes.
- How can we leverage AI with KGs?
- How can quality issues in big data (inconsistency and incompleteness) be addressed and integrated in the reasoning processes?
- How can we improve the explainability of AI systems with knowledge graphs?

Acknowledgments The research the authors presented in this chapter is partly financed by the European Union (H2020 PLATOON, Pr. No: 872592; H2020 LAMBDA, Pr. No: 809965; H2020 SINERGY, Pr. No: 952140) and partly by the Ministry of Science and Technological Development of the Republic of Serbia and Science Fund of Republic of Serbia (Artemis).

References

Alvarez, E. B. (2020). Editorial: Smart data management and applications. *Special Issues on Mobility of Systems, Users, Data and Computing, Mobile Networks and Applications*.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing, 79–80*, 3–15. https://doi.org/10.1016/j.jpdc.2014.08.003.

Auer, S., Bryl, V., & Tramp, S. (2007a) *Linked open data – Creating knowledge out of interlinked data* (Vol. 8661). Springer International Publishing. https://doi.org/10.1007/978-3-319-09846-3

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., & Ives Z. (2007b). DBpedia: A nucleus for a web of open data. In Aberer K. et al. (Eds.), *The semantic web*. ISWC, ASWC 2007. Lecture notes in computer science (Vol. 4825). Berlin: Springer. https://doi.org/10.1007/978-3-540-76298-0_52.

Berners-Lee, T. (2001). The semantic web. *Scientific American, 284*, 34–43.

Berners-Lee, T. (2006). *Design issues: Linked data*. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data – The story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Dallemand, J. (2020). Smart data; How to shift from Big Data. In *How can travel companies generate better customer insights?* Retrieved from https://blog.datumize.com/smart-data-how-to-shift-from-big-data

Davenport, T. H. (2013). Analytics 3.0. Retrieved from https://hbr.org/2013/12/analytics-30

Davis, M., Allemang, D., & Coyne, R. (2004). Evaluation and market report. IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web.

Endris, K. M., Vidal, M. E., & Graux, D. (2020). Federated query processing. In V. Janev, D. Graux, H. Jabeen, & E. Sallinger (Eds.), *Knowledge graphs and big data processing. Lecture* *notes in computer science* (Vol. 12072). Cham: Springer. https://doi.org/10.1007/978-3-030-53199-7_5.

Firican, G. (2017). *The 10 vs of big data*. Retrieved from https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Fletcher, J (2019, March 6). KGCNs: Machine learning over knowledge graphs with tensor flow. *TowardsDataScience.com*. Retrieved from https://towardsdatascience.com/kgcns-machine-learning-over-knowledge-graphs-with-tensorflow-a1d3328b8f02

Ge, M., Bangui, H., & Buhnova, B. (2018). Big data for Internet of Things: A survey. *Future Generation Computer Systems, 87*, 601–614.

Jain, S. (2021). *Understanding semantics-based decision support*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781003008927.

Janev, V. (2020). Ecosystem of big data. In V. Janev, D. Graux, H. Jabeen, & E. Sallinger (Eds.), *Knowledge graphs and big data processing* (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-53199-7_1.

Janev, V., & Vraneš, S. (2009). *Semantic Web technologies: Ready for adoption?* IEEE IT Professional, September/October, 8–16. IEEE Computer Society.

Janev, V., & Vraneš, S. (2011). Applicability assessment of semantic web technologies. *Information Processing & Management, 47*, 507–517. https://doi.org/10.1016/j.ipm.2010.11.002.

Janev, V., Mijović, V., & Vraneš, S. (2018). Using the linked data approach in European e-government systems. *International Journal on Semantic Web and Information Systems, 14* (2), 27–46. https://doi.org/10.4018/IJSWIS.2018040102.

Janev, V., Paunović, D., Sallinger, E., & Graux, D. (2020). LAMBDA learning and consulting platform. In *Proceedings of 11th International Conference on eLearning*, 24–25 September 2020, Belgrade, Serbia, Belgrade Metropolitan University.

Kern, R., Kozierkiewicz, A., & Pietranik, M. (2020). The data richness estimation framework for federated data warehouse integration. *Information Sciences, 513*, 397–411. ISSN: 0020-0255. https://doi.org/10.1016/j.ins.2019.10.046.

Lakshen, G., Janev, V., & Vraneš, S. (2020). Arabic Linked Drug Dataset Consolidating and Publishing. Computer Science and Information Systems. Retrieved from http://www.comsis.org/archive.php?show=ppr751-2005

Laney, D. (2001). *3D data management: controlling data volume, velocity, and variety*. Application Delivery Strategies, Meta Group.

Liu, Y., Wang, Q., & Hai-Qiang, C. (2015). Research on it architecture of heterogeneous big data. *Journal of Applied Science and Engineering, 18*(2), 135–142.

Mami, M. N., Graux, D., Scerri, S., Jabeen, H., Auer, S., & Lehmann, S. (2019). Uniform access to multiform data lakes using semantic technologies. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services* (pp. 313–322). https://doi.org/10.1145/3366030.3366054

Manyika, J. (2011). *Big data: The next frontier for innovation, competition, and productivity*. The McKinsey Global Institute (pp. 1–137).

Mijović, V., Tomasević, N., Janev, V., Stanojević, M., & Vraneš, S. (2019). Emergency management in critical infrastructures: A complex-event-processing paradigm. *Journal of Systems Science and Systems Engineering, 28*(1), 37–62. https://doi.org/10.1007/s11518-018-5393-5.

Mishra, S., & Jain, S. (2020). Ontologies as a semantic model in IoT. *International Journal of Computers and Applications, 42*(3), 233–243.

Patel, A., & Jain, S. (2019). Present and future of semantic web technologies: A research statement. *International Journal of Computers and Applications*, 1–10.

Patel, A., Jain, S., & Shandilya, S. K. (2018). Data of semantic web as unit of knowledge. *Journal of Web Engineering, 17*(8), 647–674.

Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Shanzhi Wang, S. (2020). Machine learning methods in drug discovery. *Molecules, 25*, 5277.

Patrizio, A. (2018, December 03). IDC: Expect 175 zettabytes of data worldwide by 2025. *Network World*. https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web, 8*(3), 489–508.

Rahman, M. A., & Asyhari, A. T. (2019). The emergence of Internet of Things (IoT): Connecting anything, anywhere. *Computers, 8*, 40. https://doi.org/10.3390/computers8020040.

Sheth, A. (1997). Panel: Data semantics: What, where and how? In R. Meersman & L. Mark (Eds.), *Database applications semantics. IAICT* (pp. 601–610). Boston, MA: Springer. https://doi.org/10.1007/978-0-387-34913-826.

Thusoo, A., Borthakur, D., & Murthy, R. (2010). Data warehousing and analytics infrastructure at Facebook. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data SIGMOD 2010* (pp. 1013–1020). ACM.

Tiwari, S. M., Jain, S., Abraham, A., & Shandilya, S. (2018). Secure semantic smart HealthCare (S3HC). *Journal of Web Engineering, 17*(8), 617–646.

Wang, L. (2017). Heterogeneous data and big data analytics. *Automatic Control and Information Sciences, 3*(1), 8–15.

Woods, W. (1975). What's in a link: Foundations for semantic networks. In *Representation and understanding* (pp. 35–82).

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web – Interoperability, Usability, Applicability, 7*(1), 63–93. https://doi.org/10.3233/SW-150175

Valentina Janev is a Senior Researcher at the Mihajlo Pupin Institute, University of Belgrade, Serbia. She received the PhD degree in the field of Semantic Web technologies from the University of Belgrade, School of Electrical Engineering. Since 2006, she has taken part in many research projects funded by the European Commission (LAMBDA, SINERGY, SLIDEWIKI, LOD2, MOVECO, EMILI, GEO-KNOW, GENDERTIME, HELENA, SHARE-PSI, PACINNO, FORSEE, Web4Web and others), coordinating two of them (see LAMBDA and SINERGY). She has published 1 authored book, 1 edited book and around 90 papers as journal, book, conference, and workshop contributions in these fields. She serves as an expert evaluator of EC Framework Programme Projects; as a reviewer and an Editorial Board Member of respectable international journals; as well as a member of the Program Committees several International Conferences including ESWC, ISWC, SEMANTiCS, CENTERIS, and ICIST.

Author Queries

Chapter No.: 4 502671_1_En

Query Refs.	Details Required	Author's response
AU1	Ref. "Auer et al. 2014" is cited in text but not provided in the reference list. Please provide details in the list or delete the citation from the text.	
AU2	Please check whether the output of artwork is appropriate as presented for Fig. 4.1 as the part image seems to be blurred.	
AU3	The citation "Mami 2020" has been changed to "Mami et al. 2019" to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.	
AU4	The citation "Auer et al. 2007" has been changed to "Auer et al. 2007a, b" to match the author name/date in the reference list. Please check if the change is fine in this occurrence and modify the subsequent occurrences, if necessary.	
AU5	References "Bizer et al. (2009), Firican (2017), Patrizio (2018), Paulheim (2017)" were not cited anywhere in the text. Please provide in text citation or delete the reference from the reference list.	