Semantic Data Ingestion for Intelligent, Value-Driven Big Data Analytics

Jeremy Debattista
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
debattij@scss.tcd.ie

Judie Attard
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
attardj@scss.tcd.ie

Rob Brennan
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
rob.brennan@cs.tcd.ie

***Abstract*—In this position paper we describe a conceptual model for intelligent Big Data analytics based on both semantic and machine learning AI techniques (called AI ensembles). These processes are linked to business outcomes by explicitly modelling data value and using semantic technologies as the underlying mode for communication between the diverse processes and organisations creating AI ensembles. Furthermore, we show how data governance can direct and enhance these ensembles by providing recommendations and insights that to ensure the output generated produces the highest possible value for the organisation.**

***Index Terms*—AI ensembles, Intelligent Analytics, Semantics, Data Governance**

I. Introduction

Big data analytics for value creation are now at the forefront of digital transformation [1]. The last decade has seen AI analytics techniques such as deep learning become mainstream. Vast data resources are needed to feed these techniques, so they have grown in tandem with Big Data. However, as the expectations for AI-based analytics grow, so does the importance of applying a network of AI techniques to address business-level problems. We term these networks as *AI ensemble*. An *AI ensemble* is a network of analytics processes working together to provide decision support, predictions or insights. Each process uses specialised AI analytics to address part of a problem, e.g. image understanding or sentiment analysis, which requires the use of AI technique-specific semantic models. We envisage a need for wider and more diverse ensembles than the ones that are currently used in machine learning or data science, spanning currently fragmented AI approaches [2]. AI ensemble diversity grows with (1) increasingly complex application domains, and (2) the trend of using multiple, heterogeneous data sources (data modes) in business analytics. Ensembles are often needed to produce the vast pools of annotated data for deep learning as unsupervised rules-based or statistical techniques are used to bootstrap analysis systems [3].

This research has received funding from the Irish Research Council Government of Ireland Postdoctoral Fellowship award (GOIPD/2017/1204), the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 713567 (EDGE), the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund.

Data is becoming an indispensable commodity, and with ever-growing flow of data through various heterogenous (un/semi-structured) sources such as sensors or social media, stakeholders are being creative in analysing this data to make innovative products. Nonetheless, whilst raw data is relatively easily passed between machine learning processes within an analysis pipeline or ensemble, each process *is isolated* in terms of end-to-end understanding. Hopkins et. al [4] defines *systems of insight* as ones that "include people, process, and technology that close the loop between data, digital insight, and action through software". Unfortunately, current AI development environments have largely ignored these systems level requirements, making development costly, slow and error prone as key system features are not supported or available as reusable components. This limits the ability to share domain or analytics insights between analysis processes to create meaningful feedback loops without expensive, hard to repeat, bespoke engineering. Understanding is also key to elevating analytics processes to adaptively address business problems from the point of view of optimising the use of organisational resources (data assets, expert time) and reducing the need for human oversight (autonomous analytics). At the machine level, understanding and reasoning is enabled through machine processable models of meaning (semantics).

Unfortunately the AI field is highly fragmented and each subfield uses specialised, machine-processable semantic models. For example, the AI disciplines of knowledge engineering and natural language processing have siloed semantic modelling techniques such as knowledge graphs or word embeddings. If one specialised analytics component uses one semantic model form it is extremely difficult to enable effective communication with others. Thus, unless extremely "narrow" ensembles of individual AI techniques are used, for example stacked learning or taking the mean of similar model outputs, diverse teams of domain and AI experts are needed to build and integrate multi-stage, diverse analysis pipelines to solve business problems. An alternative ensemble engineering approach, common in data science, is flattening the data and analytic process outputs to a lowest common denominator format for exchange, e.g. CSV files or JSON stores, thus losing the context and meaning behind the data. A common consequence of applying ensembles using data from

different sources (organisations or processes), is the loss of traceability or governance semantics (e.g. policies, consents) associated with data. Therefore, this create limitations on cross-organisation AI analytics.

In this position paper, we present our concept of building value-driven AI analytics ensembles, encompassing heterogenous data sources from diverse organisations and processes. Our presumption is based on the coordination of activities, including the understanding of data collaboration, understanding of data sources, usage and value creation of data as well as discovery and establishing trust [5]. This requirement for a common understanding and coordination can be generalised as a need for AI ensemble governance which establishes the required structures, guidelines, processes and tools to support complex activities involved in optimising the AI-data value chain, e.g. for data collaboration, understanding, discovery and trust. Hence, the aim is to build these AI ensembles methods and tools to support collaborative systems of analytic processes aligned to a business goal that are capable of efficiently dealing with high volume, velocity and variety data.

The rest of the paper is structured as follows. First we will discuss a motivation for our proposed idea in Section II. In Section III we will provide the current state of the art in this field of research, followed by the presentation of an architectural outline of the processes we propose in Section IV. We will present our final remarks in Section V.

II. Motivation: Realising UN Sustainable Goals using Data

In 2015, the United Nations (UN) adopted 17 goals[1] to achieve the Sustainable Development Agenda by 2030. In this paper we discuss goal number 11[2] - Sustainable Cities and Communities, to position and motivate our concepts. This goal looks at the challenges with maintaining cities in a way that jobs and prosperity thrives and grows, however, without fatally impacting land and resources. The UN predicts [6] that 95% of urban expansion will take place during the next decades in developing countries, and this is to the detriment of the rising population (currently stands at 828 million people) living in slum areas. Furthermore, whilst cities occupy just 3% of the Earths land, urban energy contributed to 60-80% energy consumption and around 75% carbon emission. This continuous growth without any addressing will negatively affect humans and could also result into lower gross domestic product (GDP) and life expectancies. Therefore, as the UN has put it in their document Sustainable Cities: Why They Matter [6], it is of utmost importance to build "urban resilience [is crucial] to avoid human, social and economic losses".

The targets set by the UN require diverse expertise and data on a multitude of specific subjects, from geospatial information to deforestation and census statistics and even previous law cases to ensure successful implementation of the goals 10 identified targets. The use of analytics and geospatial data is already tried and tested in Ireland[3], nonetheless, making sense out of big heterogeneous data to come up with the best strategy to tackle targets using the minimal effort possible requires a lot of coordination between different organisations having expertise to collect, analyse and interpret the different kinds of data. In our conceptual idea, the AI analytics ensembles will help stakeholders (mainly local, regional, and national authorities) efficiently collect and monitor the UN designed metrics. Taking a holistic approach ensures that gathered data from various NGOs and enterprises of heterogeneous nature and domains such as geospatial and legal can be automatically understood and aggregated with other open and crowd-sourced knowledge. Generated AI ensembles can be aimed at (a) gaining insights on how cities are configured and changed over time; (b) insights on buildings efficiencies in order to avoid potential disasters (natural or not); (c) provide information on cities infrastructure for predictive maintenance; and (d) identify and monitor changes in national legislation on outside development zone areas.

III. State of the Art

In this section we describe the research context and technological progress, focusing on three major aspects: AI analytics for multimodal understanding (AI analytics ensembles), data governance, and knowledge interoperability and re-use.

A. Analytics for multimodal understanding

Research in multimodal data understanding started about two decades ago for information retrieval applications. These applications necessitated the use of multi-stage analytic ensembles. Domain specific solutions were devised, for instance for video broadcasters, to tackle sport video analysis, summarisation and archiving [7]. Standard feature extraction techniques from image, video and audio processing fields were then fed into machine learning techniques (e.g. Support Vector Machines, Adaboost, Random Forests) to automatically tag the streams with relevant keywords to facilitate retrieval. More recent approaches such as deep learning harmoniously integrate feature extraction and machine learning capabilities with neural networks, leveraging media-specific expert knowledge to independently analyse multiple data streams. Recurrent objects, actions or events can now easily be learnt provided that enough labelled data is available to fine tune the AI. For instance, the ever-growing amount of social media data is being used to efficiently develop AI capabilities, by enabling the automatic captioning of these materials [8]. When collecting and labelling data is not suitable for training deep learning models, the most recent trend is to use a simulated environment for creating virtual digital spaces for the AI to interact and learn from. Fusing sensor data with new or traditional data sources like document stores or social media channels is an open area of research [9] that requires common understanding between the analysis techniques used. Natural language understanding is a key technology for leveraging much of this content in an analytics ensemble.

[1] https://www.un.org/sustainabledevelopment/

[2] https://www.un.org/sustainabledevelopment/cities/

[3] http://irelandsdg.geohive.ie/

B. Data Governance

The Data Management Association (DMA) defines data governance as "the exercise of authority and control (planning, monitoring and enforcement) over the management of data assets" [10]. This emphasis on centralised control of structured records does not match the reality of today's heterogeneous, networked, federated, multi-modal data sources. Moreover, data governance is particularly relevant when data is segregated into silos, e.g. providing added value by tracking data provenance. Historically, Weill and Rosss organisational approach to data governance [11] has dominated; however it focuses on roles and responsibilities rather than information system architectures, interfaces, events or algorithms. The DMAs view is more concrete and defines processes, roles and formal goals for better decision-making, assuring compliance, increasing efficiency and business integration. Al-Ruithe et al. [12] emphasise the importance of monitoring and measuring tools to support data governance. Yet, Brous et al. state "evidence is scant as to which data governance processes should be implemented, what data governance should be coordinating, or how data governance could be coordinated" [13]. Thus, there is an opportunity for new technological approaches to data governance, especially evidence-based approaches.

Data value is recognised as a "key issue in information systems management" [13]. Nonetheless, Viscusi et al. [14] recently reconfirmed Moody and Walshs earlier assertion that there is no consensus on how to measure information value.

C. Knowledge Interoperability and Management

Knowledge interoperability or data understanding requires access to many datasets with diverse syntax, semantics and access technologies. Two key technologies for automating access, cataloging and ingestion of arbitrary data are Ontology Based Data Access (OBDA) and knowledge extraction for dataset understanding. OBDA [15] aims to provide high-level access to (usually relational) data sources, which might be very large and with a complex structure. We refer to these data sources as our "data pools". The high-level access is achieved by providing a conceptual layer in the form of an ontology that defines a shared vocabulary, models the application domain, hides the structure of the data sources, and enriches incomplete data with background knowledge, therefore resulting in "semantic data pools". The ontology is connected to the data sources through a declarative specification given in terms of mappings that relate symbols in the ontology (classes and properties) to (SQL) views over the data. The W3C standard R2RML [16] was created precisely with the goal of providing a language for such mappings. The ontology, together with the mappings, exposes a high-level conceptual view of the underlying data in terms of a virtual RDF graph, which users can query using the SPARQL query language, without the need to understand the data sources, the relation between them, or the encoding of the data. Among the state-of-the-art systems supporting the virtual OBDA approach are Ontop [17] and D2RQ [18] amongst others. Note that the traditional relational OBDA framework is currently not able to support NoSQL systems because they may have non-first normal-form tables and views.

Automated analysis and understanding of data resources for applications like data cataloging and training data recommender systems relies on rich, interoperable dataset descriptions and supervised and unsupervised techniques for structured data transformation, co-reference identification or interlinking with knowledge graphs, entity recognition, relationship extraction and mapping or ontology learning . In order to efficiently managing datasets within the semantic data pools, dataset metadata needs to be extracted from a variety of formats and uniformly represented in a way that supports inference. Similarly dataset contents can be extracted to an ontology for machine understanding of the data . Dataset metadata, including provenance and data quality, will are often represented in a uniform way using standards and well known vocabularies such as DataID [19], Prov-O [20] and DQV [21]. Typically a combination of knowledge extraction from structured sources and information extraction techniques for natural language must be used. Although, especially for natural language, deep learning extraction methods have outperformed statistical or rules-based approaches the overhead in labelling is often considered disproportionate. Hence "data programming" approaches to learning labelling from user supplied rules is gaining popularity, especially for noisy data [22].

IV. From Big Raw Data to Meaningful Analytics

In order to realise our aim, we need to define a **knowledge graph-based AI ensemble engineering meta-model** that describes AI analytics techniques, a domain specific language for semantic mediators, AI ensembles, data governance, data value, data assets, and data value chains. This meta-model will enable us to create methods, tools and reusable components for developing, deploying and optimising AI analytics ensembles for extreme-scale data-intensive systems dealing with distributed multi-modal data. In this section we will discuss the four major components, namely (a) the RDF-based metamodel, (b) the AI analytics ensemble toolkit, (c) the Semantic data pool stack, and (d) the value-driven AI ensemble governance, that need to be implemented in order to create a holistic platform to enable the definition, construction and deployment of AI ensembles.

A. Metamodel for AI Analytics Ensembles

The solution will be driven by a RDF-based meta-model (Figure 1) derived from state of the art models for data engineering, analytics and data governance, mainly drawing upon W3C and ISO standards. These models will provide a common basis for controlled and interoperable information exchange within the data-intensive analytics development, deployment and governance processes among the diverse tools to be developed and extended within the solution. Since they are self-descriptive, machine readable knowledge graphs rather than simple data models, they will assist and improve software development tools, service management functions and analytics components understanding of the data, technologies and

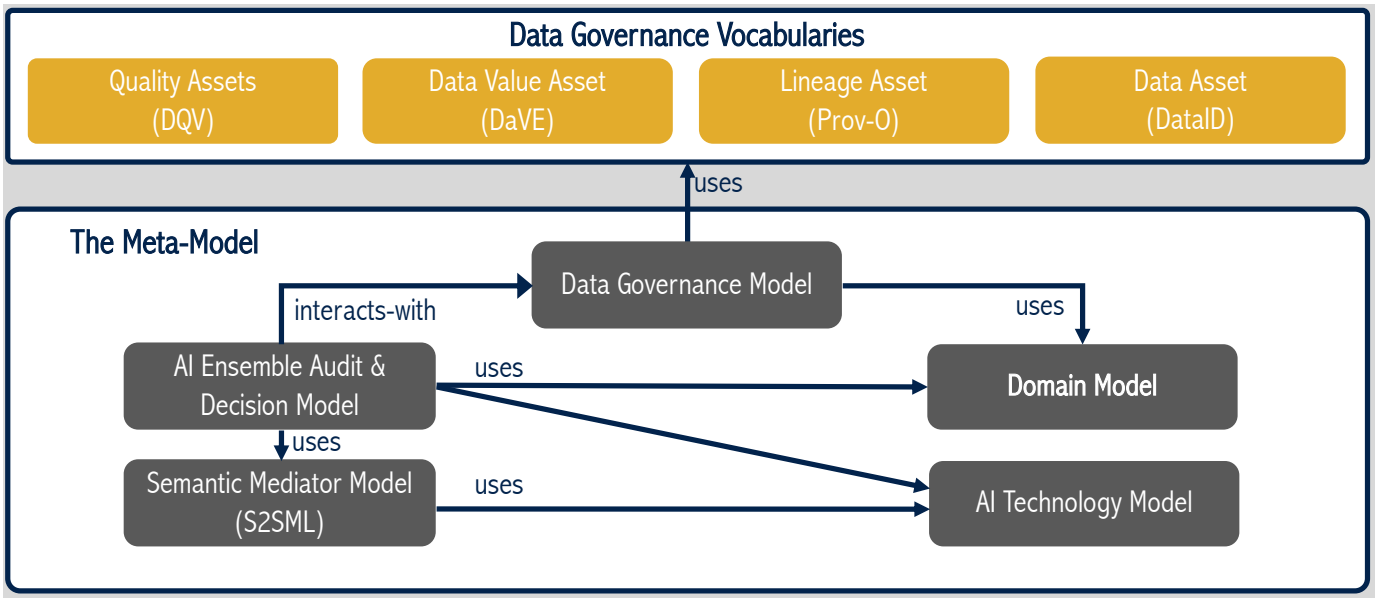

Fig. 1. AI Analytic Ensemble Metamodel

the business problem context in order to increase flexibility and improve decision-making. The meta-model is divided into five interrelated sub-models: the *semantic mediator model*, *AI technology model*, *domain model*, *data governance model*, and the *AI ensemble audit & decision model*. These models will be either developed from scratch, or where possible reused and extended following ontology engineering best practices.

The *semantic mediator model* will be used to define concepts and vocabulary for a domain specific language, the semantic to semantic model mapping language, S2SML. It will extend the pattern of the W3C Relational database to RDF mapping language, R2RML [16], by identifying key concepts and attributes for mapping between the semantic models used by AI analytics techniques.

The *AI technology model* will capture domain knowledge about AI analytics for an ensemble. It will document concepts that enable the definition of the properties (e.g. parameters and structure) of AI libraries, tools and technologies. It is anticipated that the AI technology model will be able to reuse elements of the ALIGNED software engineering life-cycle model [23] and build upon AI documentation from sources such as the human-oriented machine learning and AI catalogues like Algorithmia[4].

The *domain model* will define the common concepts that best describe specific application domains as defined by the domain's knowledge expert.

The *AI ensemble audit and decision model* will describe concepts and properties related to the building and configuration of ensembles, their data sources and value chains and will be used to support a decision and explanation service. It will leverage the data governance model. It will document configuration and deployment for AI analytics ensembles.

[4]https://algorithmia.com/algorithms

Finally a *data governance model* should be build upon a process reference model for data governance and data quality management (MAMD2.0) [24]. This will connect four assets that produces metadata in a semantic format: Quality, Data Value, Lineage, Data. The W3C provenance (PROV) model standard [20] will be used as a basis for specifying activities, agents and entities in the data governance meta-model. This will enable interoperability with standard PROV services such as meta-data repositories based on PROV AQ (access and query) and wider enterprise workflow and information integration applications. The W3C data quality vocabulary (DQV) standard [21] will be used to describe datasets quality, whilst data value vocabulary (DaVE) [25] will act as basis for describing the data value metrics and dimensions. The DataID [19] will be used as a metadata specification to describe data assets.

B. AI Analytics Ensemble Toolkit

The focus of the toolkit (Figure 2) is to enable the construction of business-value driven, governed AI analytics ensembles based on the third party AI libraries, models or tools. These are becoming increasingly prevalent, for example through marketplaces in the IBM Watson[5] or open source AI platforms such as H20.ai[6]. Thus the central process within this toolkit will be to create a knowledge-driven AI Ensemble Builder which will be command line tool/service suitable for integration into IDEs such as Jupyter or script-based data-science oriented analytics development environments. It will use the domain model, analytics business goals (encoded via the data governance platform), an innovative **AI knowledge graph** and the data broker (provided by the data governance

[5]https://www.ibm.com/watson/
[6]www.h20.ai/

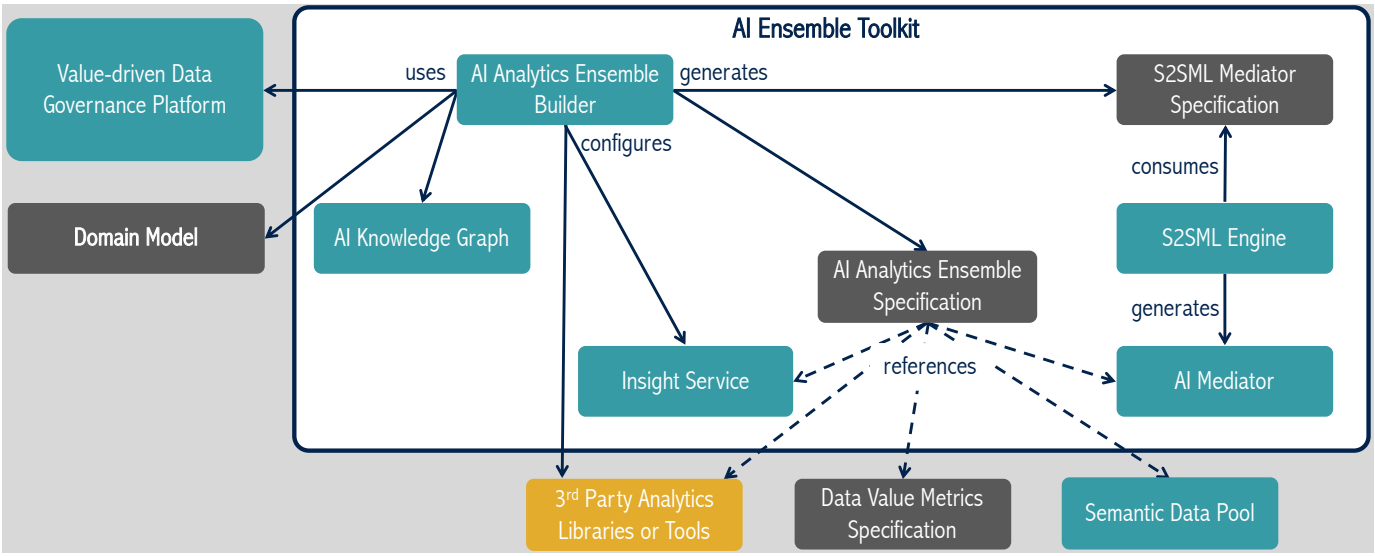

Fig. 2. AI Ensemble Toolkit

platform) to generate recommendations for datasets and analytics components needed for an ensemble to satisfy the business goals. Once a particular configuration is selected it will generate an AI Analytics Ensemble configuration and the set of **S2SML mediator specifications** to connect the analytics. The **S2SML Engine** service which is capable of generating mediator components to connect the AI Analytics will also be developed. This includes support for coreference identification, concept and instance transformation, exchange and querying.

In order to support the AI Analytics Ensemble Builder a **new open knowledge graph** for AI Analytics as a focus point for machine-readable records about AI tools, models and configurations. This can be bootstrapped by leveraging and improve upon tools such as MEX [26] and creating (semi-)automated, federated knowledge extractors for human-readable AI data, techniques descriptions, pseudocode and even open source ML and AI code.

Domain models will be used along with the data governance platforms data broker by the AI Analytics Ensemble Builder to identify which datasets might be suitable for a particular task. The data broker will also support making recommendations for appropriate training data or even pre-trained models. All of these decisions and configuration details will be captured in an instance of the ensemble audit model.

The AI Ensemble Builder will produce a set of semantic model to semantic model mapping language (S2SML) specifications for each connection in the Ensemble. These domain specific language specification will be consumed by the **S2SML Engine** which will be able of producing mediation components such as entity co-reference services, dataset transformers and query translators. The idea is to provide a unified framework for describing semantic model to semantic model mapping, drawing upon current approaches like rdf2vec [27] and supporting AI analytics tool-chains based upon them.

The insight service, together with the data broker, will identify and store high value data assets or analytics outputs. An internal knowledge graph will be used for storage and interlinking of analytics results with business purpose (domain model), the relevant value creation model, and the ensemble AI components configuration parameters. User feedback on the AI ensemble will be incorporated within this knowledge graph to ensure that the AI analytics and data governance processes are adjusted for better analytics at subsequent executions. This will enable the AI analytics components to support longer term learning and feedback loops.

C. Semantic Data Pool Stack

The increase of streaming data from heterogeneous sources and the introduction of scalable data pipeline services (such as AWS Data Pipeline [7] and Apache Beam [8]) gave enterprises the opportunity to transform their traditional centralised data warehouses into data pools (also known as data lakes). This is done by fusing their proprietary data with selected external data in order to ensure that they provide a better and potentially more personalised service to the end user. Nonetheless, this newly discovered data is still centralised in a way to the enterprise maintaining it as when this is used, the external source data is wrangled into their proprietary format, which makes it challenging for enterprises who want to share their data to perform some common tasks. We assume that enterprises will still keep their proprietary data pool, therefore, we will provide a semantic bridge where data within different entities can be used within our AI analytics ensembles on the cloud in order to utilise multi-modal data sources (e.g. sensors, documents, and enterprise data) and heterogeneous data representations like relational databases, JSON or hypertext corpora. Therefore, the focus of this module is a high-performance semantic data pool

[7]https://aws.amazon.com/datapipeline/
[8]https://beam.apache.org/

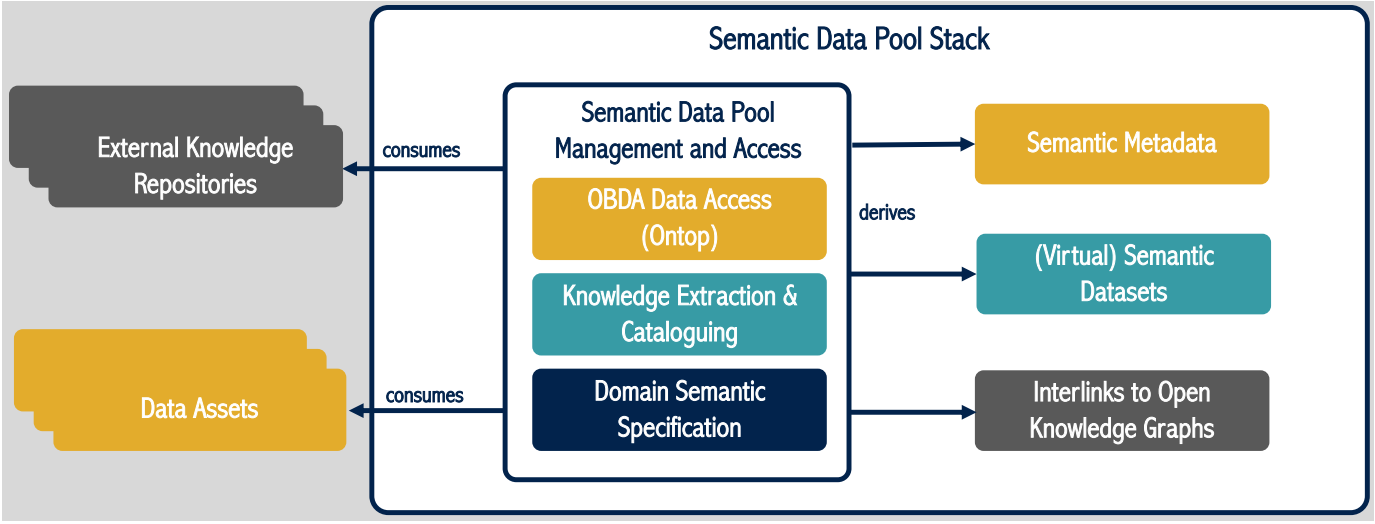

Fig. 3. Semantic Data Pools

infrastructure that directly tackles the large data wrangling overhead faced in analytics deployments and establishes a syntax and data model-neutral way for machine understanding and access to heterogeneous, federated data pools.

The semantic data pool stack (Figure 3) will have an underlying semantic schema that describes the domain of the enterprise based on the following components:

1) **Ontology-Based Data Access (OBDA)** - The mature Ontop platform, will be used to provide high performance query-based data access across a wide variety of formats. Ontop has to be extended to support non-relational data sources (e.g. CSV files) whilst making sure that queries are optimised to ensure fast execution.
2) **Knowledge extraction, Interlinking and Cataloguing** In order to efficiently discriminate between and process diverse data pools, AI analytics ensembles need access to accurate metadata and links to global knowledge graphs to provide context for each data pool. In this process, tools for automated dataset domain identification, metadata extraction using the DataID suite of vocabularies, ontology learning to support OBDA and interlinking, fusion and ingestion by the DBpedia [28] open knowledge graph will be catered for. This enables semantic search and easy integration of catalogs in the data broker.
3) **Data Pool Management** In order to manage this raw big data, a general infrastructure for distributed access to data pools with the associated storage and computational resources ahas to be developed. This will also integrate the previously mentioned components into a stack of semantic tools.

D. Value-driven AI Ensemble Data Governance

Figure 4 illustrates the Data Governance process. At its core, the solution could be developed upon an off-the-shelf data governance platforms (e.g Collibra's DGC[9]), leveraging on the available workflow support, policy engine, and polished user interfaces supporting all data governance roles. Extensions and processes could be built around the data governance platform to serve the main aim of the proposed concept. The main technical input to the governance platform are the virtual semantic datasets and metadata provided by the Semantic Data Pool Stack. These dataset descriptions will be captured by the governance platform using common Linked Data serialisation formats such as JSON-LD or RDF/XML. These datasets are crawled by a metadata harvester and this information is registered and certified directly with a new semantic data discovery sandbox. The discovery sandbox will act as an intelligent data broker, where apart from being a super catalogue of catalogues (describing all of the connected semantic data pools), it will also provide a value-driven data recommender system used to match a data asset to a data consumer (e.g. an AI analytics process) based on a domain-specific data value assessment. The recommender system will use datasets metadata defined by the data governance meta models to make an informed recommendation. This requires the sourcing of scalable quality assessment and value monitoring services in order to be able to derive these aspects from the various datasets that can be extracted from the different enterprise semantic data pools. Following the quality assessment methodology described in Debattista et al. [29], we will extend Luzzu with the appropriate quality metrics which will serve as the data quality service. This will be automatically invoked whenever a new dataset (or a new version of an existing dataset) is discovered by the governance platform. The results of Luzzu[10] are consumed by the platform

[9]https://www.collibra.com/data-governance-solutions/data-governance-center

[10]Metadata is generated using the Dataset Quality Vocabulary (daQ) [30] which is semantically equivalent to the DQV model as expected by the data governance model

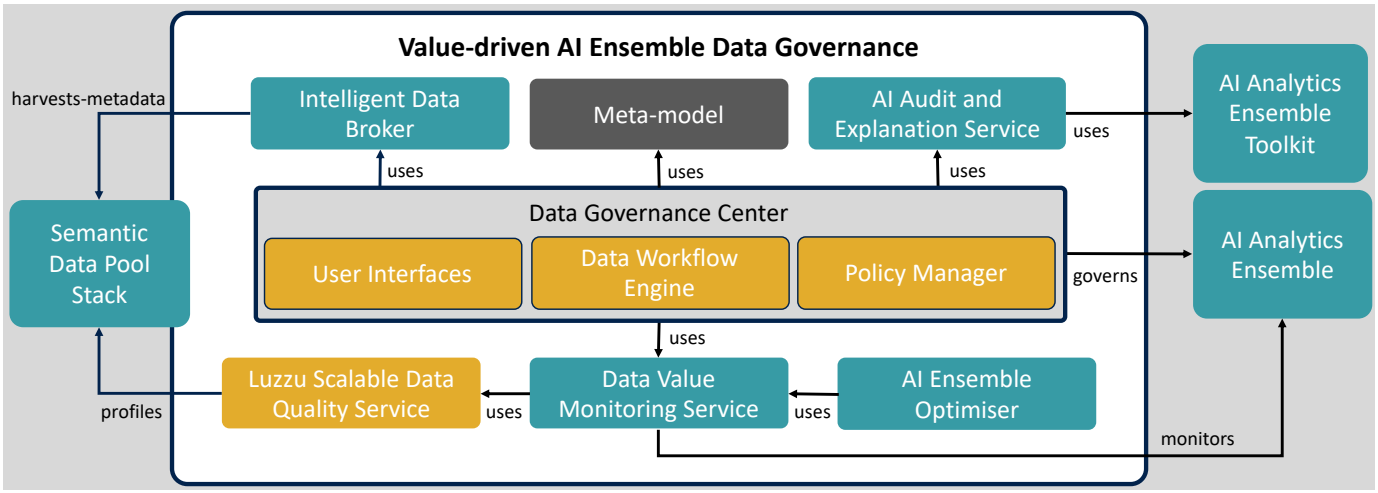

Fig. 4. Value-driven Data Governance Platform

for further processing. With regard to value monitoring service, Brennan et al. [31] describe a capability maturity model for data value monitoring. This will be prototyped as a new data value monitoring service based on dataset and ensemble-specific metrics and the results, defined using DaVe, the Data Value ontology [25], will also be consumed by the governance platform.

Specification of data value chains will be achieved in the AI analytics ensemble optimisation tool. Data value chains are not static, and thus the data value monitoring service will provide recommendations to adapt the enterprises data value chain to ensure that the enterprise benefits from maximal productivity efficiency and ultimately maximum economic impact. Data governance platforms might also ensure that any recommendation given is in line with the policies defined by the owners of the semantic data pool infrastructure. Such considerations are handled by the policy manager.

One of the known limitations of data-driven AI is the opaque nature of results. Our reference model, based on MAMD 2.0, will alleviate this limitation as we services will be built to provide some of the assurances and explanations required by a modern regulatory environment, such as GDPR, for business adoption. By consuming all the required information, the governance platform would be in a position to explain the stages of a value chain, trace lineage, shows the data quality and link to the business domain model in order to demonstrate results in terms of data value. This process will include ontological inference based on the domain model, insight service and externally linked knowledge graphs to provide the business context behind analytics recommendations.

V. Final Remarks

In this position paper we have presented a conceptual model for AI ensembles. More specifically, we discussed how raw data from diverse heterogenous sources can be unified and governed using an interoperable method in order enable value-driven intelligent analytics. In Section IV we described four important components that holistically will enable us to achieve the aim of building AI ensembles methods and tools to support collaborative systems of analytic processes aligned to a business goal that are capable of efficiently dealing with high volume, velocity and variety data, by supporting the transfer of meaning between heterogeneous analytics processes.

Acknowledgement

We would like to thank Giovanni Schiuma, Markus Helfurt, Pieter De Leenheer, Eamonn Clinton, Diego Calvanese, Christian Dirschl, Ismael Caballero, Hans Viehmann, and Rico Richter for their valuable insights and comments on this work.

References

[1] T. Davenport and J. Harris, "Competing on analytics," 2017.
[2] W. Knight, "AIś language problem," https://www.technologyreview.com/s/602094/ais-language-problem/, last Accessed: 2018-05-15.
[3] V. A. Krylov, E. Kenny, and R. Dahyot, "Automatic discovery and geotagging of objects from street view imagery." *CoRR*, vol. abs/1708.08417, 2017.
[4] B. Hopkins, M. Goetz, and N. Yuhanna, "The anatomy of a system of insight," 2017.
[5] P. D. Leenheer, "Data governance in a big data era," 2017.
[6] UN, "Sustainable cities: Why they matter," 2016.
[7] A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan, "Browsing sports video: trends in sports-related indexing and retrieval work," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 47–58, March 2006.
[8] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions." *CoRR*, vol. abs/1412.2306, 2014.
[9] A. Bulbul and R. Dahyot, "Social media based 3D visual popularity." *Computers & Graphics*, vol. 63, pp. 28–36, 2017.
[10] *The DMA Guide to the Data Management Body of Knowledge, AUTHOR = Data Management Association, YEAR = 2010, Technics Publications LLC, 2010,*.
[11] P. Weill and J. W. Ross, *IT Governance : How Top Performers Manage IT Decision Rights for Superior Results*. Boston, Mass.: Harvard Business School Pr., 2004.
[12] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, "Key dimensions for cloud data governance." in *FiCloud*, M. Younas, I. Awan, and W. Seah, Eds. IEEE Computer Society, 2016, pp. 379–386.
[13] P. Brous, M. Janssen, and R. Vilminko-Heikkinen, "Coordinating decision-making in data management activities: A systematic review of data governance principles." in *EGOV*, ser. Lecture Notes in Computer Science, vol. 9820. Springer, 2016, pp. 115–125.

[14] G. Viscusi and C. Batini, "Digital information asset evaluation: Characteristics and dimensions," in *Smart Organizations and Smart Artifacts*, L. Caporarello, B. Di Martino, and M. Martinez, Eds. Cham: Springer International Publishing, 2014, pp. 77–86.

[15] A. Poggi, D. Lembo, D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," *Journal on Data Semantics*, vol. 10, pp. 133–173, 2008.

[16] S. Das, S. Sundara, and R. Cyganiak, "R2rml: rdb to rdf mapping language," World Wide Web Consortium, Tech. Rep., 2012. [Online]. Available: https://www.w3.org/TR/r2rml/

[17] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao, "Ontop: Answering sparql queries over relational databases." *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2017.

[18] C. Bizer and A. Seaborne, "D2RQ - treating non-RDF databases as virtual RDF graphs," in *ISWC2004 (posters)*, November 2004.

[19] M. Freudenberg, M. Brmmer, J. Rcknagel, R. Ulrich, T. Eckart, D. Kontokostas, and S. Hellmann, "The metadata ecosystem of DataID." in *MTSR*, ser. Communications in Computer and Information Science, vol. 672, 2016, pp. 317–332.

[20] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, "PROV-O: The PROV ontology," World Wide Web Consortium (W3C), W3C Recommendation, 2013.

[21] R. Albertoni, A. Isaac, C. Guéret, J. Debattista, D. Lee, N. Mihindukulasooriya, and A. Zaveri, "Data quality vocabulary (DQV)," World Wide Web Consortium (W3C), W3C Interest Group Note, June 2015.

[22] A. J. Ratner, C. D. Sa, S. Wu, D. Selsam, and C. R, "Data programming: Creating large training sets, quickly." in *NIPS*, 2016, pp. 3567–3575.

[23] M. Solanki, B. Bozic, M. Freudenberg, D. Kontokostas, C. Dirschl, and R. Brennan, "Enabling combined software and data engineering at web-scale: The aligned suite of ontologies." in *International Semantic Web Conference (2)*, ser. Lecture Notes in Computer Science, vol. 9982, 2016.

[24] A. G. Carretero, F. Gualo, I. Caballero, and M. Piattini, "MAMD 2.0: Environment for data quality processes implantation based on iso 8000-6x and iso/iec 33000." *Computer Standards & Interfaces*, vol. 54, pp. 139–151, 2017.

[25] J. Attard and R. Brennan, "A semantic data value vocabulary supporting data value assessment and measurement integration." in *ICEIS (2)*. SciTePress, 2018, pp. 133–144.

[26] D. Esteves, P. N. Mendes, D. Moussallem, J. C. Duarte, A. Zaveri, and J. Lehmann, "Mex interfaces: Automating machine learning metadata generation." in *SEMANTICS*, A. Fensel, A. Zaveri, S. Hellmann, and T. Pellegrini, Eds. ACM, 2016, pp. 17–24. [Online]. Available: http://dblp.uni-trier.de/db/conf/i-semantics/semantics2016.html#EstevesMMDZL16

[27] P. Ristoski and H. Paulheim, "RDF2Vec: rdf graph embeddings for data mining." in *International Semantic Web Conference (1)*, ser. Lecture Notes in Computer Science, vol. 9981, 2016, pp. 498–514.

[28] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web Journal*, 2014.

[29] J. Debattista, S. Auer, and C. Lange, "Luzzu – a methodology and framework for Linked Data quality assessment," *Data and Information Quality*, vol. 8, no. 1, Oct. 2016.

[30] J. Debattista, C. Lange, and S. Auer, "Representing dataset quality metadata using multi-dimensional views," in *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, 2014, pp. 92–99.

[31] R. Brennan, J. Attard, and M. Helfert, "Management of data value chains, a value monitoring capability maturity model." in *ICEIS (2)*. SciTePress, 2018, pp. 573–584.