A Conceptual Architecture for AI-based Big Data Analysis and Visualization Supporting Metagenomics Research

Thoralf Reis, Thomas Krause,
Marco X. Bornschlegl, and Matthias L. Hemmje

University of Hagen, Faculty of Mathematics and Computer Science, 58097 Hagen, Germany {thoralf.reis, thomas.krause, marco-xaver.bornschlegl, matthias.hemmje}@fernuni-hagen.de

Abstract. This paper targets to introduce an architecture for Artificial Intelligence (AI) based Big Data Analysis and Visualization supported metagenomics research based on the AI2VIS4BigData Reference Model. Metagenomics research covers the examination of huge amounts of data to improve the understanding of microbial communities. Technological and methodical improvements in Big Data Analysis drive progress in metagenomics research and thereby support practical applications like, e.g., the analysis of cattle rumen with the research goal of reducing the negative impact of cattle breeding on global warming. AI2VIS4BigData is a reference model for the combined application areas of Big Data Analysis, AI, and Visualization. Its purpose is to support scientific and industrial activities with guidelines and a common terminology to enable efficient exchange of knowledge and information and thereby prevent "reinventing the wheel". The general applicability of the AI2VIS4BigData model for metagenomics has been validated in a previous publication. As a next step, this paper derives a conceptual architecture that specifies a possible adaption of AI2VIS4BigData for metagenomics. For this, three new metagenomic publications utilizing AI and Visualizations are assessed.

Keywords: Metagenomics · Big Data · AI · Visualization · AI2VIS4BigData.

1 Introduction and Motivation

Metagenomics research analyzes relationships within whole microbial communities while genomics research focuses on the analysis of genes or the genome of a single organism [1]. A practical example for metagenomics research is the investigation of the rumen microbiota regarding its influence in cattle greenhouse gas emissions and food conversion efficiency [2] as cattle are a major contributor to climate change and relevant for food security, two significant challenges society is facing [2]. The demand for data in metagenomics research is significantly bigger than for regular genomics research: the investigation of relationships and coherence between organisms or genes in and between metagenomic samples

[1] requires biological researchers to process, store, and exchange big amounts of data e.g. via specialized bioinformatics databases [3]. Hence, metagenomics research benefits on a large scale from progress and development in Big Data Analysis such as decreasing costs for storage and processing of huge amounts of data. With the EU-funded MetaPlat[1] project, scientists from different research institutions with either Big Data Analysis or bioinformatics background worked together to develop the MetaPlat platform. This cloud based Big Data Analysis platform is specialized to analyze metagenomics data like, e.g., rumen microbiota [2]. For an effective analysis of Big Data, the platform empowers the researchers to utilize cutting-edge technology such as Artificial Intelligence (AI) [2] and various forms of Information Visualization (IVIS) [4] to provide the researchers with visual feedback of their activities and enable them to identify new insights.

To define the vague term Big Data, a popular approach is to follow the data management challenges outlined by Doug Laney [5]. These challenges comprise three dimensions (the three v's): variety (ambiguous data manifestations regarding e.g. data format, data structure or data semantics), volume (big amount of data), and velocity (high frequent data inflow) [5]. By this definition, the sheer volume of data in metagenomics research allows labeling it as Big Data. The collective term AI summarizes techniques and methods such as symbolic AI or Machine Learning (ML) to implement intelligence for machines (in contrary to human or animal natural intelligence) [6]. Example application scenarios of AI in metagenomics research of rumen are the analysis of data through clustering [3] or the training of classifiers to categorize data samples [2]. Big Data and AI are closely connected to each other [6]: Big Data is very useful to derive, validate, apply, and enhance AI models while AI-driven algorithms enable the exploration of Big Data and its potential. Visualization of data, processing steps as well as AI model development is an important link between both application areas. It enhances comprehension and decreases entry barriers for new users. In addition, visualization offers the chance to meet the growing demand for explainability and transparency of AI

With [7], the AI2VIS4BigData Reference Model for research and practical applications in the application areas Big Data Analysis, AI, and Visualization was introduced. Its objective is to provide a common specification as well as a common basis for discussion and thereby reduce the risk of inefficiency through reinventing the wheel and solving problems that have already been solved elsewhere. The reference model's theoretical applicability was evaluated in an expert round table workshop featuring presentations from three practical application domains: health care, economics, and metagenomics [8]. Until now, the reference model was validated only for one metagenomics research application [9]. This paper targets to validate three further metagenomics research applications from the MetaPlat project to assess if an architecture can be derived.

Within the remainder of this paper, the AI2VIS4BigData Reference Model and the three assessed metagenomics research publications from the MetaPlat project are introduced, the pursued architecture modeling approach is presented

[1] https://metaplat.eu

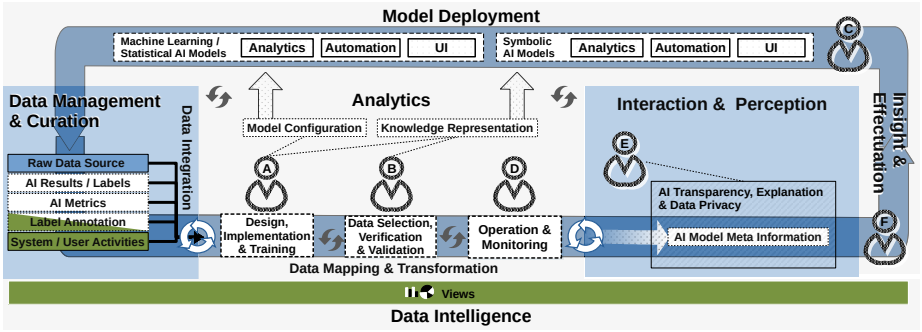

Fig. 1. AI2VIS4BigData - A Reference Model for AI Supporting Big Data Analysis

(Section 1.3), a multi-layered conceptual architecture is introduced in Section 2 together with an initial validation in Section 3 before this paper concludes with outlining its contributions and providing an outlook (Section 4).

1.1 AI2VIS4BigData Reference Model

The AI2VIS4BigData Reference Model (Figure 1) was derived through projecting the AI lifecycle phases of AIGO's AI System Lifecycle [10] onto the IVIS4BigData Reference Model [11] considering the different AI models (ML or statistical AI models as well as symbolic AI models) for supporting Big Data Analysis, AI data, and AI user stereotypes [7]. It contains the three processing steps *Data Management & Curation*, *Analytics*, *Interaction and Perception* accompanied with a data intelligence layer for user interaction and User Interfaces (UI) of IVIS4BigData [11], a reference model that target to *"close the gap in research with regard to information visualization challenges of Big Data Analysis as well as context awareness"* [11]. AI2VIS4BigData introduces a model deployment layer that spreads over the three processing steps [7]. AI models are executed directly within the data and information loop which links the deployed models to the input data they need for execution and compute output data that is fed back into the Big Data Analysis system [7]. The remaining activities of AI system life cycle phases are displayed within the analytics layer as *Design, Implementation & Training*, *Data Selection, Verification & Validation* as well as *Operation & Monitoring* and interconnected through bidirectional arrows emphasizing the iterative nature of AI model design [7]. The different reference model elements are linked to five clearly distinguished AI user stereotypes (model designer, domain expert, model deployment engineer, model operator, model end user, and model governance officer) and four clearly distinguished Big Data Analysis user stereotypes (system owner, data scientists, management consultants as well as directors including C-levels) [7].

1.2 Assessed Metagenomic Use Cases

In [9] a conceptual workflow for metagenomic studies was presented and demonstrated using two previously published metagenomic use cases. The first of these

use cases [12] was the visualization of gene dependencies using a whole-genome approach and a new framework for improved correlation measurement between genes. The second publication [13] analyzed the relationship between microbiomes in feces and rumen using a taxonomic analysis of partial genome sequences (barcode sequences). Together they cover the two main branches of metagenomic analyses (taxonomic and functional). It was shown in [9] that the metagenomics analysis workflow extracted from these publications can be mapped directly onto the AI2VIS4BigData Reference Model therefore validated its relevance for the field of metagenomics.

This section will introduce three additional publications in metagenomics research that serve as a base to further validate and transform this conceptual workflow into a generic architecture. The publications were selected as they are practical examples for metagenomic analysis (which can represent Big Data Analysis applications depending on the sample size), carried out by different researchers and most importantly, they describe the usage of statistical methods or ML as well as Visualization. Although all selected publications originate from the MetaPlat project, they represent different research approaches like, e.g., the analysis of genes or the analysis of OTUs. In addition, the homogeneous MetaPlat terminology eases the architecture derivation.

The publication *A Metagenomics Analysis of Rumen Microbiome* [2] by P. Walsh et al. demonstrates a metagenomic analysis of the "Bos taurus" rumen microbiome using ML models in a cloud based environment. For optimal performance and scalability, a queueing system is used between individual components, thus enabling asynchronous and parallel execution. After importing raw sequence data into the system, it is written to one of these processing queues which feed into a similar metagenomic analysis workflow that uses the QIIME toolset to perform data cleanup and clustering of sequences into Operational Taxonomic Units (OTUs). The workflow assigns taxonomic labels to these OTUs. In an analytics step, various ML models are used to classify the samples into phenotypes using the taxonomic data of the previous steps as an input. Finally, the publication showcases various visualizations ranging from a taxonomic composition chart to plots of algorithmic accuracy and other AI metrics.

In *Analysis of Rumen Microbial Community in Cattle through the Integration of Metagenomic and Network-based Approaches* [3], H. Wang et al. functionally analyze the rumen microbial community in cattle through application of a network-based approach: the authors construct a co-abundance network utilizing the *"relative abundance of 1570 microbial genes"* [3] that enables them to identify functional modules. In doing so, they present a method to automatically determine a cutoff threshold value to generate the co-abundance network in the first place [3]. While the first publication [2] uses partial sequences sufficient to identify and analyse the taxonomic composition, this publication is based on whole genome data which enables the analysis of genes. Together they cover the two main branches of metagenomic studies. To construct the co-abundance network used in the publication, the short reads generated by next-generation sequencing platforms are assembled into longer sequences. These sequences are

then matched to the KEGG[2] database to identify genes (and associated metadata) present in the samples. Using the relative abundances of these genes, correlations can then be computed by analyzing how the abundance of one gene affects the abundance of other genes across the various samples. Since the presence or absence of a correlation is not always distinguishable from statistical noise, a suitable cutoff value is then determined using an automated computational method. Using the cutoff values, a network graph can be constructed that represents genes as nodes and the correlation strength as the length of edges connecting these nodes.

As third and last assessed publication, M. Wang et al., the authors of *Understanding the relationships between rumen microbiome genes and metabolites to be used for prediction of cattle phenotypes* [4] combined metabolomics with metagenomics in order to identify differences in diets and methane emissions from rumen metabolites and microbial genes. They analyzed 36 rumen samples and identified the difference in the response of rumen microbes to different basal diets which down the road affect cattle methane emissions [4]. The study starts from gene abundance data of cattle rumen obtained from previous studies on the experiment designed by Roehe et al. [14]. The abundance data was cleansed and transformed before conducting multiple activities to determine correlations between genes and metabolites related to the differences in diets in the experiment design. The correlation data was then used to build correlation networks as well as various other plots and result tables.

1.3 Dicsussion, Conclusion, and Identification of remaining Architectural Challenges

In order to arrive at a generic architecture that enables the management, analysis, and visualization of metagenomic data as well as the fusion with other health related data and knowledge, the first step is mapping the introduced metagenomics publications to the generic stages of the AI2VIS4BigData Reference Model ("Data Management & Curation", "Analytics" and "Interaction & Curation"). This is easy to validate as all three publications include steps to ingest, manage or cleanup metagenomic sequences, all of them include statistical or ML methods for analytics and also all of them produce one or more visualizations. The same was previously already demonstrated in [9]. Therefore, it is proposed that an architectural model should explicitly model these stages.

Looking at the papers in detail, further requirements for a comprehensive architectural model can be derived: The first publication describes the importance of using individual components that communicate through asynchronous mechanisms like, e.g., queuing systems to achieve high performance and scalability. The impact of Big Data and ML in Metagenomics is also mentioned as a challenge in [9]. A suitable architecture should therefore aim to separate individual parts and components of the system where possible so that they can operate and scale individually. The second publication [3] shows the need of additional

[2] Kyoto Encyclopedia of Genes and Genomes, https://kegg.jp

knowledge sources like, e.g., gene databases for the analysis of metagenomic sequences. Our proposed architecture should therefore support the ingestion and persistence of these additional data sources into a knowledge network that can be used by metagenomic workflows. The third publication [4] is important as it does not start from raw sequence data but from intermediate results obtained from other studies. Our architecture should be able to reuse the same intermediate results for several distinct analyses thus requiring the persistence of these intermediate results. This requirement also partially addresses the challenge of "Reproducibility" mentioned in [9] and the area of "AI Transparency, Explanation & Data Privacy" of the AI2VIS4BigData model. All three publications differ significantly in the exact steps executed in the analysis phase and the visualizations produced. It is therefore important that the analysis is done in a modular fashion where the order and type of steps is dynamic and that a wide range of visualizations is supported.

2 AI2VIS4BigData Conceptual Architecture supporting Metagenomics Research

This paper introduces the AI2VIS4BigData architecture (Figure 2) for processing and analysis of metagenomic data in an AI and Big Data environment. It was designed by extending the Big Data Analysis and Visualization architecture of IVIS4BigData [11] with AI and metagenomic aspects in order to fulfill the metagenomics requirements outlined in Section 1.3. The architecture is vertically split into three pillars separating the components for metagenomics data integration and processing (domain-specific input), AI and data science modeling and configuration (AI analysis input) from the components responsible for result visualization and data generation (output). This is based on the design principle of Separation of Concerns (SoC) [15] and makes it easier to develop, scale or exchange the components separately. Each of these three pillars is structured into three layers following the Model View Controller (MVC) pattern [16] with a shared persistence layer interconnecting all three pillars. The bottom layer represents the model, the top layers represent the view while the middle layers contain the controllers. Metagenomics-specific architecture elements are a dedicated user, knowledge and data artifacts within the input layer, assets and knowledge networks in persistence layer as well as domain-specific end user interfaces. The following rough description of the individual layers and components follows the flow of data, starting from the top left at data input and ending with result visualization at the top right corner:

Knowledge & Data Input. Within this layer, expert users or systems ingest metagenomics-related knowledge and data into the system. These information comprise biological and genetical knowledge (e.g. protein metadata or knowledge automatically extracted from scientific publications) as well as diagnostic and subject data (e.g. metagenomic sequences).

AI Integration & Fusion. This layer contains all services and methods to integrate the various domain-specific inputs into the system, to perform a data fusion and persist it as structured content or knowledge network. The se-

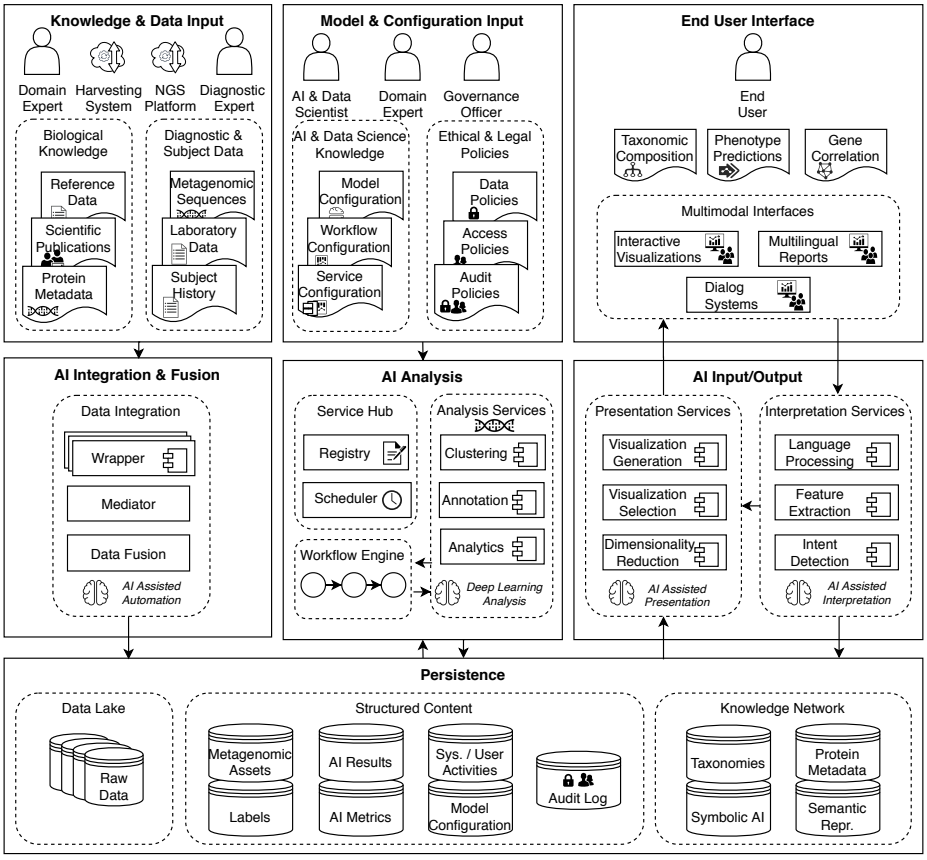

Fig. 2. AI2VIS4BigData Conceptual Architecture Supporting Metagenomics Research

mantic integration is realized through implementation of the mediator wrapper approach.

Model & Configuration Input. The necessary knowledge and information for configuring the AI applications within the system is provided by AI and data science expert users within this layer. The input contains the required knowledge to register and schedule all AI services and to select appropriate analysis methods and algorithms. The additional AI2VIS4BigData role of the *Governance Officer* ensures legal compliance and maintaining ethical standards through providing relevant constraints.

AI Analysis. The middle layer is responsible for performing analysis tasks on behalf of the user. A workflow system together with a service registry allow for flexible configuration of the required analysis steps while the scheduler manages the execution of these steps on distributed or local computing nodes. Intermediate and final results are stored persistently.

Persistence. The persistence layer targets to store various types of data and enable data exchange between overlying layers. Raw data is stored in a data lake with little to none processing performed to improve reproducibility and trans-

parency of the system. Structured data includes parsed genetic sequences, intermediate results from analysis processes and other kind of schema-bound data. Lastly a knowledge network tries to represent biological and medical knowledge as well as semantic rules required for Symbolic AI in a machine readable way.

AI Input/Output. The purpose of this layer is to intelligently interpret the intentions of the system's end user (e.g. through applying natural language processing) and present the information that is relevant for them in a suiting form (e.g. after performing a dimensionality reduction or selecting appropriate visualization techniques).

End User Interface. The end user interface layer contains the multimodal interfaces through which the system's end users access its data and information. These interfaces comprise visualizations, reports and dialogue systems that present the domain-specific artifacts (e.g. taxonomic compositions).

3 Initial Validation and Remaining Challenges

The proposed architecture specifies all areas of the AI2VIS4BigData Reference Model. The area "Data Management & Curation" of the reference model is addressed by the left pillar. The "Analysis" area is covered by the second pillar and especially the "AI Analysis" layer. Finally, the "Interaction & Perception" area is implemented through the right pillar. The architecture also implements all requirements that were outlined previously. A detailed mapping of the requirements to the architecture elements would be beyond the scope of this paper, yet is planned for future work. Individual components for input, analysis and visualization are strictly split by the three pillars and communicate asynchronously through the persistence layer allowing for flexible scaling and Big Data processing. Additional knowledge sources are supported by providing a data agnostic input layer together with a mediator wrapper architecture for data integration. The persistence layer ensures that reproducibility and transparency is possible by storing intermediate and final results. Finally, a flexible workflow system and a service registry support the heterogeneity of metagenomic studies and allow easy integration of new analysis methods. The remaining challenges for the architecture comprise a harmonization with the IVIS4BigData architecture, a generalization for application domains beyond metagenomics research, a technical specification as well as a proof of concept technical implementation. Since the selected publications were limited to the MetaPlat project, the assessment of practical applicability for the introduced architecture in metagenomics research beyond MetaPlat is a further remaining challenge.

4 Conclusion and Outlook

In the course of this paper, three MetaPlat publications were assessed that analyze rumen microbiota through metagenomics research utilizing Big Data Analysis, AI as well as visualization. Objective of this assessment was the derivation of a AI2VIS4BigData-based conceptual architecture for real-life application in this three-fold research area. The resulting AI2VIS4BigData conceptual architecture supporting metagenomics research was introduced in Section 2. It consists of

seven layers arranged alongside the three levels of the MVC pattern. As outlook, future work is planned to overcome the challenges introduced in Section 3.

References

1. F. Engel, M. Fuchs, P. M. Kevitt, M. Hemmje, and P. Walsh, "A Metagenomic Content and Knowledge Management Ecosystem Platform," 2019.
2. P. Walsh, C. Palu, B. Kelly, B. Lawor, J. T. Wassan, H. Zheng, and H. Wang, "A Metagenomics Analysis of Rumen Microbiome," *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 2077–2082, 2017.
3. H. Wang, H. Zheng, F. Browne, R. Roehe, R. J. Dewhurst, F. Engel, M. Hemmje, and P. Walsh, "Analysis of Rumen Microbial Community in Cattle through the Integration of Metagenomic and Network-based Approaches," *2016 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 198–203, 2017.
4. M. Wang, H. Zheng, H. Wang, R. J. Dewhurst, and R. Roehe, "Understanding the relationships between rumen microbiome genes and metabolites to be used for prediction of cattle phenotypes," in *BIBE 2019; The Third International Conference on Biological Information and Biomedical Engineering*. VDE, 2019, pp. 1–5.
5. D. Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," META Group, Tech. Rep., 2001.
6. ISO, "ISO/IEC JTC 1/SC 42 Artificial Intelligence," 2018. [Online]. Available: https://isotc.iso.org/livelink/livelink/open/jtc1sc42
7. T. Reis, M. X. Bornschlegl, and M. L. Hemmje, "Towards a Reference Model for Artificial Intelligence Supporting Big Data Analysis," *To appear in: Proceedings of the 2020 International Conference on Data Science (ICDATA'20)*, 2020.
8. ——, "AI2VIS4BigData: Qualitative Evaluation of a Big Data Analysis, AI, and Visualization Reference Model," *To appear in: Lecture Notes in Computer Science*, vol. LNCS 10084, 2020.
9. T. Krause, B. Andrade, H. Afli, H. Wang, H. Zheng, and M. Hemmje, "Understanding the Role of (Advanced) MachineLearning in Metagenomic Workflows," *To appear in: Lecture Notes in Computer Science*, vol. LNCS 10084, 2020.
10. OECD, *Artificial Intelligence in Society*, 2019.
11. M. X. Bornschlegl, "Advanced Visual Interfaces Supporting Distributed Cloud-Based Big Data Analysis," Dissertation, University of Hagen, 2019.
12. H. Zheng, H. Wang, R. Dewhurst, and R. Roehe, "Improving the Inference of Co-occurrence Networks in the Bovine Rumen Microbiome," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2018.
13. B. G. N. Andrade, F. A. Bressani, R. R. C. Cuadrat, P. C. Tizioto, P. S. N. de Oliveira, G. B. Mourão, L. L. Coutinho, J. M. Reecy, J. E. Koltes, P. Walsh, A. Berndt, L. C. A., J. C. P. Palhares, and L. C. A. Regitano, "The structure of microbial populations in Nelore GIT reveals inter-dependency of methanogens in feces and rumen," *Journal of animal science and biotechnology*, vol. 11, p. 6, 2020.
14. R. Roehe, R. J. Dewhurst, C.-A. Duthie, J. A. Rooke, N. McKain, D. W. Ross, J. J. Hyslop, A. Waterhouse, T. C. Freeman, M. Watson, and R. J. Wallace, "Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance," *PLOS Genetics*, pp. 1–20, 2016.
15. E. W. Dijkstra, "On the role of scientific thought," in *Selected writings on computing: a personal perspective*. Springer, 1982, pp. 60–66.
16. M. Fowler, *Patterns of enterprise application architecture*. Addison-Wesley Longman Publishing Co., Inc., 2002.