A Reference Architecture for Smart Digital Platform for Personalized Prevention and Patient Management

Amal Elgammal[1,2(✉)] and Bernd J. Krämer[1]

[1] Scientific Academy for Service Technology e.V. (ServTech), Potsdam, Germany
{amal,kraemer}@servtech.info
[2] Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

Abstract. The maturity of a new generation of information technologies, including the internet of things (IoT), wearables, cloud computing, Artificial Intelligence (AI) and machine learning, has led to the advent of smart domains, such as smart manufacturing, smart logistics, and smart healthcare. Smart healthcare brings unlimited opportunities to solve many of the problems of traditional medical systems, with the ultimate goal of realizing 4P medicine (Predictive, Preventive, Personalized, Participative). However, to realize this ambitious vision in such a highly regulated multi-disciplinary and sensitive domain, a mine of challenges needs to be effectively and efficiently addressed. A smart health digital platform that integrates all relevant (semi-) structured and unstructured health-related data is fundamental. The platform should incorporate a variety of care data, including vital medical information from medical records, current medication, imaging studies, lifestyle, genetic, demographic, psychological & psychosocial and patient-provided health data from exercise or health monitoring applications and medical pathways. These will lead to improving post-operative planning, reduce medical risks and costs, and generate more accurate therapy and increased Quality of Life (QoL) for patients. The main contribution of this article is a reference architecture for a smart digital platform for personalized prevention and patient management that acts as a roadmap for further R&D in this domain.

1 Introduction

Traditional healthcare is being revolutionized as a result of the maturity and synergy of a new generation of information technologies, including IoT, wearables, big data, cloud computing, simulations, AI, and machine learning. The concept "Smart Healthcare" was born out of the concept of "Smart Planet" proposed by IBM in 2009. Smart healthcare is a health service system that uses technology to dynamically access information, connect people, materials, and institutions related to healthcare, and then actively manages and intelligently responds to medical ecosystem needs [1, 2]. Smart healthcare

This research is partially funded by the EC Horizon 2020 project QUALITOP, under contract number H2020 - SC1-DTH-01-2019 – 875171.

M. Aiello et al. (Eds.): Papazoglou Festschrift, LNCS 12521, pp. 88–99, 2021.
https://doi.org/10.1007/978-3-030-73203-5_7

opens unlimited opportunities to solve many of the problems of traditional medical systems. This includes: (i) promoting the collaboration between all involved stakeholders, including patients, relatives, doctors, nurses, healthcare providers, healthcare institutions, public health, scientists and policymakers, (ii) improving the monitoring of patient activities outside the traditional care setting – including medication adherence management, chronic disease management support and other interventions – and (iii) reducing costs through improved care coordination and operational improvements founded on the increased visibility of patient activities – reducing unnecessary service utilization, and allocating resources more efficiently.

To realize the real potential of smart healthcare, many challenges are continuously emerging. In particular, smart healthcare lacks macro guidance, which results in poorly planned development goals and ultimately a waste of resources [1]. Furthermore, smart healthcare inherits the problems of Big data 3Vs (Volume, Velocity, and Variety) [2]. In these sensitive and highly regulated domains, the situation is even aggravated by special requirements to data protection, security, and privacy regulations as in GDPR[1]. Particularly, in healthcare, big data challenges are compounded by the fragmentation and dispersion of heterogeneous data among various stakeholders. Besides, for any successful implementation of a smart healthcare system, it should be founded on an agile, robust, reliable, secured, and scalable platform by considering healthcare data standards and information exchange standards. On top of such a platform, various querying, simulations, and data analytics functionalities are enabled and can be adapted to the requirements of the respective stakeholders.

The main contribution of this article is a reference architecture for a smart digital platform for personalized prevention and patient management meeting these requirements. The reference architecture also acts as a roadmap for further R&D activities in this direction and it realizes 4P medicine [3].

The rest of this paper is structured as follows: Sect. 2 analyses related work and identifies the gaps in smart healthcare platforms described in the literature. Based on these findings, Sect. 3 presents the proposed reference architecture for a smart digital platform for personalized prevention and patient management and discusses its main components. Finally, conclusions and future work directions are drawn in Sect. 4.

2 Related Work

Research efforts that specifically focus on smart healthcare platforms/architectures are rare in the scientific literature that are discussed in Sect. 2.1. Other related parallel streams of work addressing specific challenges of individual components of a smart healthcare platform are discussed thereafter in the following sub-sections. They include: (i) big data integration & interoperability techniques, (ii) Domain-Specific Languages (DSL) for data-intensive applications, and (iii) data security and privacy.

2.1 Smart Healthcare Architectures

Prominent work in this direction is reported in [4–7]. The authors in [4] propose an IoT-aware, smart architecture for automatic monitoring and tracking of patients.

[1] General Data Protection Regulation: https://gdpr.eu, last access: 03.08.2020.

The architecture integrates with a smart hospital system, which relies on RFID, WSN, and smart mobile gadgets interoperating with each other through a low-power wireless area network. Similarly, the work in [5] builds an IoT-aware architecture for smart healthcare coaching systems that assists families and patients in their daily living activities and hence improving the QoL of patients by allowing them to get the medical care they need at home. Analogously, the study in [6] proposes an IoT network management system architecture that is reliable, effective, well-performing, secure, and profitable for caregivers. Finally, a survey of a 5G-Based smart healthcare network is reported in [7] concluding with a range of challenges and future R&D directions.

The architectures discussed in these papers ignore the challenge of big data variety (big data integration & interoperability) that represents one of the major challenges in big data, especially in the healthcare domain. This lack limits the validity and applicability of the proposed approaches in practice. Furthermore, these approaches offer no solution for managing large scale data processing systems; they neither address data security and privacy nor investigate the role of DSLs for data-intensive applications. The proposed reference architecture described in Sect. 3 considers all these vital points and builds on the concept of modern data architecture as opposed to the traditional concept of data warehouse (cf. Sect.3.2). These concepts form the foundation of employing advanced analytics mechanisms including runtime monitoring, predictive analytics, and simulation for informed clinical decision-making. In addition, they ensure the agility, scalability, security, and privacy of the smart healthcare platform.

2.2 Big Data Integration and Interoperability Techniques

Big data in healthcare refers to electronic health data sets so large and complex that they are difficult to manage with traditional data management tools and methods not only because of their volume but also because of the diversity of data types and the speed at which they must be managed [8]. In healthcare, big data challenges are compounded by the fragmentation and dispersion of data among various stakeholders.

Much of the focus on Big Data integration has been on the problem of processing very large sources, extracting information from multiple, possibly conflicting data sources, and reconciling the values and providing unified access to data residing in multiple, autonomous data sources. The work has mainly focused on isolated aspects of data source management relying on schema mapping and semantic integration of different sources, and its final goal was not reasoning about the content and quality of sources [9–11]. Moreover, most of that work has focused on sources from a specific domain and does not present results for largely heterogeneous sources. Most solutions follow the traditional model in which all data are loaded into a warehouse. This centralized approach is not applicable in the medical domain because hospitals and other stakeholders would never store their patient data in an external data silo.

2.3 Domain-Specific Languages for Data-Intensive Applications

Traditionally, computing models such as MapReduce [12] were proposed to specifically support data-intensive applications. Although such models suited massive-scale data processing, they permit limited application logic complexity [13]. Domain-Specific

Languages (DSLs) can be employed to circumvent such problems. DSLs are usually concise offering a set of pre-defined abstractions to represent concepts from the application domain close to real concepts and terms familiar to the experts in the domain. DSLs can ease the implementation of analytics and machine learning algorithms with the use of high-level abstractions or reusable pieces of code that hide low-level details from lay users letting them focus on the main problem at hand.

Languages like OptiML [14] enable machine-learning algorithms to take advantage of parallelism by bridging the gap between machine learning and heterogeneous big data hardware infrastructure. OptiML is a declarative, statically-typed textual programming language, in which operations support parallel executions (using the MapReduce programming model) on heterogeneous machines. But this language lacks support for a distributed environment or executions in the cloud. ScalOps [15] is another example of a declarative, statically-typed textual DSL, intending to enable machine learning algorithms to run on a cloud computing environment and overcome the lack of iteration limitation of the traditional MapReduce programming model. To date, there is very little use of DSLs in the medical domain, while the reference architecture proposed in this paper will exploit the capabilities of DSLs.

2.4 Data Security and Privacy

Access control [16] mediates every access request to resources and data managed by a system and determines whether the request should be authorized or denied. An access control system can be considered at three different levels of control: access control policy, access control model, and access control mechanism. Access control models have emerged that break the direct relationship between subject/object by introducing new concepts such as tasks, roles, rights, responsibilities, teams, etc.). RBAC Model [17] was the most popular access control model for enforcing access control where the role (job function) is the core of privileges in such a model. Unlike other access control policies, users do not acquire the permissions directly, but they acquire them through the roles they play in the organization. Attribute-Based Access Control (ABAC) [18] overcomes the limitations of RBAC. ABAC is considered more flexible than RBAC because it can easily handle contextual attributes as access control parameters [19, 20] but is more complex from a policy review's perspective.

To address the problem of content-based access control – where queries are defined according to the base tables and, then, are rewritten by the system against the user authorized view –, Oracle has proposed a fine-grained based access control approach, the Virtual Private Database (VPD) [21]. Ensuring data confidentiality in the presence of views is an important element of research. Currently, there are several research efforts devoted to addressing issues related to enforcing access control to view based approaches (i.e., materialized or virtual views). Rosenthal et al. automatically calculate derived permissions on the data warehouse with those of the sources by extending the standard SQL grant/revoke model [22, 23]. This allows automated inference of many permissions for the warehouse to systems with redundant and derived data. Finally, in [24], the authors propose a methodology that allows controlling access to a data integration system.

3 The Reference Architecture

The discussion in the previous section revealed several gaps in the state-of-the-art in realizing the ultimate potential of smart healthcare that mandates the building of smart healthcare networks (SHN). Challenges to achieving the SHN vision include a lack of:

- Common modelling formalisms,
- Configuration and deployment strategies,
- Compliance with data protection and privacy regulations with special attention to GDPR,
- Effective data governance strategies,
- Effective meta-data management,
- Advanced knowledge representation mechanisms enabling dynamic yet controlled collaboration across healthcare systems.

As a foundational component, a smart digital platform should be carefully analysed, designed, and developed to serve the purposes/functionalities/use-cases of respective, usually diverse healthcare stakeholders. Based on the literature analysis and feedback from the medical partners in the context of the EU H2020 project QUALITOP[2], we have iteratively developed and validated reference architecture for a smart digital platform for personalized prevention and patient management. A *reference architecture* describes a software system's fundamental organization, embodied in its modules and their interrelationships. It helps achieve an understanding of specific domains and provides consistency of technology implementation for solving domain-specific applications [25]. We advocate in this article that smart healthcare architectures should be built on technology and healthcare standards, which are largely ignored in related work. The smart digital healthcare reference architecture provides a holistic consolidated view of currently dispersed healthcare data silos, enabling new innovative ecosystems, and capacitating different stakeholders, e.g., patients, clinicians, nurses, family, friends, caregivers, laboratory providers/staff, insurance entities, and public authorities, to use for the continuous monitoring, simulation and analysis of patients. All this will empower proactive health and well-being management and enable 4P medicine [3]: Predictive, Preventive, Personalized, Participative.

Figure 1 presents the medical reference architecture. As shown at the bottom of the figure, data from various sources are collected to provide patient-specific health status and Quality of Life (QoL) measures of patients.

Data can be structured, occurring in the form of electronic medical records, stored and maintained in medical institutions' databases and data warehouses; other data can be semi-structured or unstructured originating from file systems, such as questionnaires regarding lifestyle data, nutrition data, QoL data, and diagnostic imaging datasets.

Privacy and GDPR are specifically addressed and given primary attention in the architecture through the adoption of an edge computing approach and enforcing role-based access. Our edge computing approach requires all computations that could be done

[2] Monitoring multidimensional aspects of QUAlity of Life after cancer ImmunoTherapy - an Open smart digital Platform for personalized prevention and patient management: https://h2020qualitop.liris.cnrs.fr/wordpress/index.php/project/.

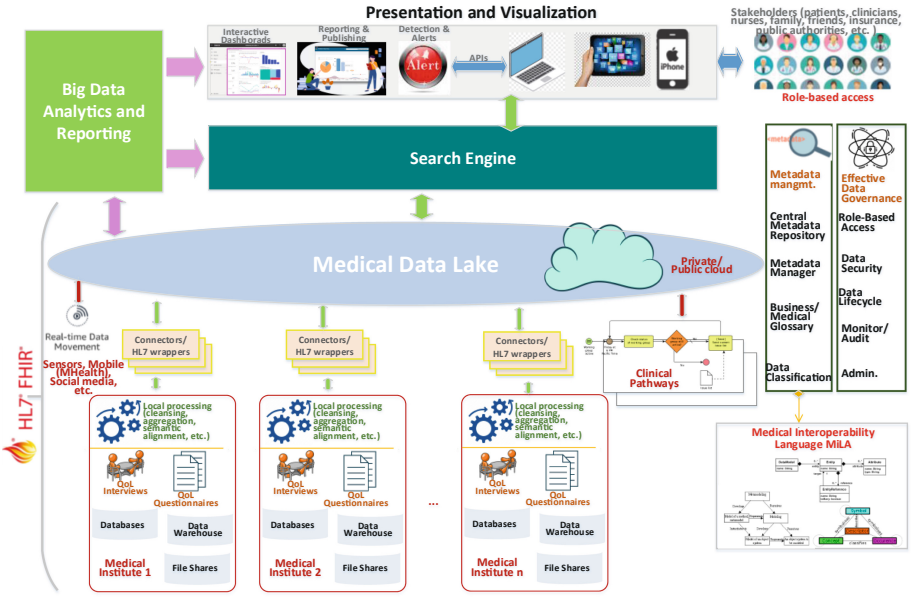

Fig. 1. A reference architecture for smart digital platform for personalized prevention and patient management

locally (without the movement/transfer of data) to be performed at the medical partners' nodes, e.g., cleansing, aggregation, semantic alignment, etc. Only then, anonymized individual or aggregated data will travel to the central data lake for decent querying and advanced analytics. *A data lake* is a centralized repository that allows the storage of all types of structured and unstructured data at any scale. Data can be stored as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions [26]. In the reference architecture presented in Fig. 1, the medical data lake is a virtual layer, which means that transferred data will not be stored in the data lake; only query/analytics results will be maintained in the data lake along with their metadata. The main components of the reference architecture are discussed in detail in the following sub-sections.

3.1 Big Data Management Layer

Due to the sensitivity and strict data protection and privacy regulations of the healthcare domain, such as GDPR, the reference architecture requires that data from different participating medical institutions will still be stored and maintained at respective local nodes (see the bottom of Fig. 1), therefore, as explained earlier in this Section an edge computing approach is adapted. The architecture supports diverse data sources and formats, including structured, semi-structured, and unstructured data, also supporting real-time data streaming from wearable sensors and devices, and runtime data emerging from clinical/medical pathways. Data at the local nodes need to go through several transformation

processes to transform raw data into smart data. The term smart data emphasizes the latent value inherent in widely dispersed and unconnected data sources. In this article, we envision the journey from raw data to smart data is:

- *Raw data -> Normalized data*: First we need to convert raw data, such as patient summary data, demographical data, results, and reports including medical images, etc. into conflict-free, homogenized data retrieved from multiple related sources that can be interpreted in a specific context.
- *Normalized data -> Contextualized data*: normalized data is then given meaning & contextual-awareness to enable orchestration & improved decision-making. For example, in the context of "Cancer assessment" relevant data include historical data, screening tests, cancer history, stage, CT scans, MRI, lifestyle and nutrition data, and others
- *Contextualized data -> Orchestrated data*: This step cross-correlates secure data across a specific domain that can be turned to actionable tasks, for example, immunotherapy treatment, at the speed of business, realizing smart data. For example, cancer assessment data can be linked with contexts such as intensive behavioural counselling, treatment of metabolic disorders, also linking a patient's primary and psychiatric care provider data, and treatments.

In the context of the EU H2020 project QUALITOP, a first big data real-life cohort of cancer patients treated with immunotherapy will be created, supporting both prospective and retrospective data coming from five EU countries participating in the project. In prospective studies, individuals are followed over time, and data about them is collected. In retrospective studies, individuals are sampled, and information is collected about their past, through interviews or questionnaires.

To ensure interoperability, enforce standard descriptions, and ensure wide applicability, information exchange standards (message-based and structured document-based) must be supported. In this context, the architecture adopts HL7-FHIR (https://www.hl7.org/fhir/), which is a powerful standard describing resources and an API for exchanging Electronic Health Records (EHRs). Therefore, the architecture integrates a set of HL7 wrappers, adaptors, and connectors that link the data management layer to the medical data lake layer in a loosely coupled manner. This enables the transfer and ingestion of needed on-demand anonymized data to the data lake for processing and advanced querying or analysis.

3.2 Medical Data Lake

The medical data lake represents the heart of the reference architecture. It is an open reservoir for the vast amount of data inherent within healthcare, which can be integrated into an analytics platform to improve decision making. The data lake employs data security and privacy mechanisms to ensure confidentiality and anonymity of data transfer to avoid misinterpretation and inappropriate conclusions by using proper annotation methodologies of the data.

A common misperception is that a data lake is a data warehouse replacement. On the contrary, a data lake is a very useful part of an early-binding data warehouse, a late-binding data warehouse, and a distributed big data processing system, such as Hadoop (https://hadoop.apache.org/). The early- binding mechanism in a data warehouse guarantees that all the data are organized and harmonized before it can be consumed. An early-binding data mechanism is not appropriate for healthcare data as it requires a lot of time to map the data before realizing value. In contrast, when a data lake with a late-binding data mechanism is employed, only the required data are organized, harmonized, and integrated instead of all the data in the data lake. This is called "schema-on-read" or "late binding" because structure and meaning are provided to the data only when the data are read (as users request).

A data lake brings value to healthcare because it stores all the data in a central repository and only maps it as needs arise. The concept of "Map-Reduce" such as in Hadoop systems, when implemented, divides the Big Data integration problem into smaller parts, assigns them to different processing nodes to solve partial tasks, and then accumulates and synthesizes the results on which it applies data-driven analytics and advanced simulation methods. A distributed big data processing system can act as a software framework to handle structured and unstructured data and host analytics mechanisms in a data lake. This approach allows data to be processed faster since the system is working with smaller batches of localized data instead of the contents of the entire warehouse. Therefore, it leads to better and faster means of high-quality response to prevent or timely address the development of new medical conditions and better knowledge for improved patient counselling. Also, it improves the patients' follow-up.

The architecture assumes that the data lake is hosted on the cloud. Given the strict data security and privacy regulations in healthcare, private clouds are recommended.

3.3 Metadata Management

Metadata and efficient metadata management capabilities are mandatory for the success of any data lake implementation, which act to simplify and automate common data management tasks. Metadata ensures that the medical data lake makes the system agile enough to accommodate and scale new types of data. Metadata gives the ability to understand lineage, quality, and lifecycle and provides needed visibility [27, 28]. Metadata is also vital because it enables data governance (cf. Sect. 3.4), the second vital component of any successful data lake implementation. In addition to technical metadata (capturing the form and structure of each data set) and operational metadata (capturing the lineage, quality, profile and provenance of data), business and medical metadata is essential to capture what the data means to stakeholders and to make data easier to find and understand. Harmonization of data digested in the data lake is done through the metadata. Therefore, the metadata management layer of the reference architecture comprises five components: (i) central metadata repository, (ii) metadata manager, and (iii) business/medical glossary that contains definitions agreed upon by stakeholders that ensures that all stakeholders have common understandings and consistently interpret the same data by a set of rules and concepts, (iv) data classification, and (v) the novel Medical Interoperability Language (MiLA), which is a DSL Interoperable Language for health status that is under development in QUALITOP. More specifically, MiLA will:

- Provide appropriate notations, constructs, and a set of operators and offer the expressive power required to integrate a wide variety of medical data sources.
- Provide constructs that make use of high-level abstraction mechanisms that can be used for analytics purposes to improve decision-making.
- Be extensible, pluggable, and parameterizable to avoid problems of current DSLs.

3.4 Effective Data Governance

Effective data governance is the second vital component (in addition to meta-data and meta-data management; cf. Sect. 3.3) for the success of any data lake implementation. Metadata enables data governance that contains the policies and standards for the management, security, quality, and use of data, enforcing data access at the enterprise level [27]. Data lake security ensures that only those users are granted access to the lake, to specific components of the system, or specific portions of the data, who own specific permissions based on the security rules defined for the data lake system [29]. The data governance strategy strictly and reliably secure three components:

- Role-based platform access and privileges: the architecture provides the components to store and process data, and therefore, security for each type or even each component should be defined and enforced. These may rely on federated identity provision, single sign-on, and SSH keys authentication.
- Network Isolation: as described in Sect. 3, the data lake may be hosted on a private cloud to prevent undesired access and protecting the data lake property. However, public clouds are a more doable option, which could be secured through VPNs and firewalls.
- Data Protection: first, data transferred from medical data sources are anonymized, second, anonymized data are encrypted while data transfer, and well as their temporary storage on the data lake for advanced processing and analysis. Once data is processed, it will be permanently deleted from the data lake. Only query and analytics results will be stored on the data lake.

3.5 Search Engine and Big Data Analytics and Reporting

On top of the data lake various search, browse, and analytics capabilities can be devised and implemented using late binding, satisfying the requirements of the various stakeholders. A key aspect of this ubiquitous collection of data is to link them together to reveal elements of the "bigger picture" and use analytics and machine learning techniques to take appropriate actions to solve the problem at hand and build a sound foundation for clinical decision-making. The combination of different sources creates a deeper understanding that leads to decision-making, action-taking, and effective problem-solving in the healthcare domain. The Search Engine allows for substantial performance improvements as well as query capabilities not supported by SQL-based engines, including faceted and text search across many data sets, advanced analytics mechanisms to employ runtime monitoring, predictive analytics, and simulation to extrapolate the course of future events from descriptive data. It is not the aim of this article to provide a review of big data analytics capabilities and use cases in healthcare. Interested readers are referred to [2] for more information about the potential of Analytics.

3.6 Presentation and Visualization

On top of the reference, architecture rests the "Presentation and Visualization" layer that takes inputs from the "Search Engine" and "Big data analytics and reporting" components, and visualize results on customized end-user interactive dashboards, supporting different platforms, e.g., laptops, medium-sized devices, and smartphones. The dashboard provides real-time visibility for clinical decision making using advanced graphic representations of event data, alarms, thresholds, KPI status, drug interactions, patients' vital signs, and performance levels.

4 Conclusions and Future Work

Smart Healthcare brings many promises to the healthcare community in solving the problems of traditional medical systems. The ultimate goal is to realize the concept of 4P medicine (Predictive, Preventive, Personalized, participative). However, to realize this ambitious vision in such a highly regulated multi-disciplinary and sensitive domain, a mine of challenges needs to be effectively and efficiently addressed. Above all, a smart health digital platform that integrates all relevant (semi-) structured and unstructured health-related (big) data is fundamental. The platform should be agile, robust, reliable, secured, and scalable that considers healthcare data standards and information exchange standards. This paper introduces a reference architecture for a smart digital platform for personalized prevention and patient management that meets these requirements and acts as a roadmap for R&D in this direction.

In the context of the EU H2020 project QUALITOP, further R&D efforts will be pursued in parallel and complementary directions to realize the building components of the reference architecture presented in this paper. Following an agile systems development approach, the prototype of an open smart digital platform for personalized prevention and patient management in Europe will be developed. By creating the first big-data real-life cohort of cancer patients treated with immunotherapy, the functional and non-functional requirements of the platform, and subsequently the proposed reference architecture will be iteratively validated and evaluated to ensure the applicability, validity, and efficacy of the proposed architecture and its analytics features.

Acknowledgment. Mike Papazoglou is one of the pioneers in service-oriented computing (SOC) and cloud computing, after having contributed ground-breaking work on database systems. He has strongly influenced the continuously growing SOC community he helped create as a co-founder of the International Conference on Service-Oriented Computing (ICSOC), which will celebrate its 30th anniversary in 2022.

Mike was also instrumental in writing the successful Horizon 2020 proposal QUALITOP whose first results are presented in this paper. Finally, Mike is not only an admirable colleague but also a long-term friend.

Bernd got to know Mike in 1983 during an EU ESPRIT project meeting at the University of Patras. Soon after this meeting, Mike joined GMD - Forschungszentrum Informationstechnik GmbH, a major German research institution for applied mathematics and computer science Bernd was also affiliated with at that time. Amal was a Ph.D. student of Mike who looks back to a fruitful scientific apprenticeship in Mike's lab from 2008 to 2012; He has always been the mentor; role

model and continuous supporter and she is honored to have him as an empowering friend since then.

We are deeply honored for the opportunity to contribute to Mike's Festschrift celebrating his 65th birthday and transition to a life with much freestyle and lesser compulsory. We wish him all the best, many more birthdays, and a state of health that does not require him to rely on our smart health platform.

References

1. Tian, S., et al.: Smart healthcare: making medical care more intelligent. Global Health J. **3**, 62–65 (2019)
2. Galetsi, P., Katsaliaki, K.: A review of the literature on big data analytics in healthcare. J. Oper. Res. Soc. **71**, 1511–1529 (2020)
3. Flores, M., et al.: P4 medicine: how systems medicine will transform the healthcare sector and society. Pers. Med. **10**, 565–576 (2013)
4. Catarinucci, L., et al.: An IoT-aware architecture for smart healthcare systems. IEEE Internet Things J. **2**, 515–526 (2015)
5. Amato, A., Coronato, A.: An IoT-aware architecture for smart healthcare coaching systems. In: 2017 IEEE 31st AINA, pp. 1027–1034 (2017)
6. Sallabi, F., Shuaib, K.: Internet of things network management system architecture for smart healthcare. In: 2016 6th DICTAP, pp. 165–170 (2016)
7. Ahad, A., et al.: 5G-based smart healthcare network: architecture, taxonomy, challenges and future research directions. IEEE Access **7**, 100747–100762 (2019)
8. Frost & Sullivan: Drowning in big data? Reducing information technology complexities and costs for healthcare organizations (2015)
9. Cafarella, M.J., Halevy, A., Khoussainova, N.: Data integration for the relational web. Proc. VLDB Endow. **2**, 1090–1101 (2009)
10. Venetis, P., et al.: Recovering semantics of tables on the web. Proc. VLDB Endow. **4**, 528–538 (2011)
11. Hassanzadeh, O., et al.: Discovering linkage points over web data. Proc. VLDB Endow. **6**, 445–456 (2013)
12. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**, 107–113 (2008)
13. Kalavri, V., et al.: m2r2: a framework for results materialization and reuse in high-level dataflow systems for big data. In: 2013 IEEE 16th CSE, pp. 894–901 (2013)
14. Chapelle, O., Li, L.: An empirical evaluation of Thompson sampling. Presented at the Proceedings of the 24th NIPS, Granada, Spain (2011)
15. Xindong, W., et al.: Knowledge engineering with big data. IEEE Intell. Syst. **30**, 46–55 (2015)
16. De Capitani di Vimercati, S., Samarati, P., Jajodia, S.: Policies, models, and languages for access control. In: Bhalla, S. (ed.) DNIS 2005. LNCS, vol. 3433, pp. 225–237. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31970-2_18
17. Ferraiolo, D.F., et al.: Proposed NIST standard for role-based access control. ACM Trans. Inf. Syst. Secur. **4**, 224–274 (2001)
18. Wang, L., Wijesekera, D., Jajodia, S.: A logic-based framework for attribute based access control. In: The 2004 ACM, FMSE, USA (2004)
19. Bertino, E., Catania, B., Damiani, M.L., Perlasca, P.: GEO-RBAC: a spatially aware RBAC. In: The 10th SACMAT, Sweden (2005)
20. Rajpoot, Q.M., Jensen, C.D., Krishnan, R.: Integrating Attributes into Role-Based Access Control. Cham, pp. 242–249 (2015)

21. Huey, P.: Using oracle virtual private database to control data access. In: Oracle Database Security Guide, Chapter 7 (2012)
22. Rosenthal, A., Sciore, E.: View security as the basis for data warehouse security. In: 2nd DMDW 2000, Sweden (2000)
23. Rosenthal, A., Sciore, E.: Administering permissions for distributed data: factoring and automated inference. In: IFIP TC11/WG11.3, Canada, pp. 91–104 (2001)
24. Haddad, M., Stevovic, J., Chiasera, A., Velegrakis, Y., Hacid, M.-S.: Access control for data integration in presence of data dependencies. In: Bhowmick, S.S., Dyreson, C.E., Jensen, C.S., Lee, M.L., Muliantara, A., Thalheim, B. (eds.) DASFAA 2014. LNCS, vol. 8422, pp. 203–217. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05813-9_14
25. Angelov, S., Trienekens, J., Kusters, R.: Software reference architectures - exploring their usage and design in practice. In: Drira, K. (eds.) Software Architecture, pp. 17–24. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39031-9_2
26. Campbell, C.: Top Five Differences between Data Warehouses and Data Lakes. Blue-Granite.com (2017)
27. Gidley, S.: Tips for managing metadata in a data lake (2017). https://www.oreilly.com/content/tips-for-managing-metadata-in-a-data-lake/
28. Sawadogo, P.N., Scholly, É., Favre, C., Ferey, É., Loudcher, S., Darmont, J.: Metadata systems for data lakes: models and features. In: Welzer, T., et al. (eds.) ADBIS 2019. CCIS, vol. 1064, pp. 440–451. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30278-8_43
29. Maroto, C.: Data Lake Security: Four Key Areas to Consider When Securing Your Data Lake. https://www.searchtechnologies.com/blog/data-lake-security