Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things – SEA4DQ'22 Report

Phu H. Nguyen, Sagar Sen
SINTEF
Oslo, Norway
{phu.nguyen, sagar.sen}@sintef.no

Beatriz Bretones Cassoli, Nicolas Jourdan
Technische Universität Darmstadt,
Darmstadt, Germany
{b.cassoli, n.jourdan}@ptw.tu-darmstadt.de

Maria Chiara Magnanini
Politecnico di Milano,
Milan, Italy
mariachiara.magnanini@polimi.it

Mikel Armendia
Tekniker
Eibar, Spain
mikel.armendia@tekniker.es

DOI: 10.1145/3573074.3573103
http://doi.org/10.1145/3573074.3573103

ABSTRACT

Cyber-physical systems (CPS)/Internet of Things (IoT) are omnipresent in many industrial sectors and application domains in which the quality of the data acquired and used for decision support is a common factor. Because of things like sensor failures and defects brought on by working in harsh and unreliable conditions, data quality might suffer. *How can software engineering and artificial intelligence (AI) help manage and tame data quality issues in CPS/IoT?* Data quality is of paramount importance for CPS/IoT. This workshop series stemmed from the common interest in data quality of the Zero-Defect Manufacturing (ZDM) Research and Innovation projects under the Horizon 2020 Framework Programme such as InterQ (https://interq-project.eu/) and DAT4.Zero (https://dat4zero.eu/). Not only for ZDM, but also in general, emerging trends in software engineering need to take data quality management seriously as CPS/IoT are increasingly data-centric in their approach to acquiring and processing data along the edge-fog-cloud continuum. This workshop provides researchers and practitioners a forum for exchanging ideas, experiences, understanding of the problems, visions for the future, and promising solutions to the problems in data quality in CPS/IoT. SEA4DQ 2022 took place on November 17th, 2022 and collocated with the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC / FSE) 2022 in Singapore. The workshop featured two great keynotes, six excellent presentations, and concluded on a high note with an extensive panel discussion.

Categories and Subject Descriptors

CCS Concepts: • **Software and its engineering → Embedded software**; *Layered systems*; • **Information systems → Database utilities and tools**; *Data compression*; *Data encryption*; **Information lifecycle management**; **Data analytics; Online analytical processing; Process control systems; Computing platforms;** • **Computer systems organization → Sensors and actuators; Embedded software**; *Sensor networks*.

General Terms

Algorithms, Management, Measurement, Performance, Design, Reliability, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Verification.

Keywords

Data Quality, Software Engineering, IoT, CPS, Industry 4.0, AI, Machine Learning, Smart Manufacturing, ZDM.

1. INTRODUCTION

Cyber-physical systems (CPS)/Internet of Things (IoT) have been omnipresent in many industrial sectors and application domains to acquire sensor data from the physical world and make decisions in real-time and based on historical analysis of long-term data. Manufacturing companies have utilized CPS/IoT to collect high frequency data from machine tools like Computer Numerical Control (CNC) machines for predictive maintenance to produce goods with zero faults. Smart healthcare systems have been used to acquire data from body sensors for health monitoring. Semi-autonomous and networked cars in the automobile industry have employed real-time sensor data for traffic management and safe navigation. To implement smart meters and reduce our household carbon footprint, the energy sector has deployed CPS/IoT to collect data on energy production and consumption.

A common aspect across industrial sectors that supports our confidence in and reliance on CPS/IoT systems is the caliber of the data collected and used for decision support. There exist many classifications for data quality in the literature [1], [2]. For instance, data quality can be categorized in dimensions such as: *data completeness* - the percentage of missing data values, *data accuracy* - data values are correct and stored in a consistent and unambiguous form, *data consistency* - refers to when same data kept at different places do not match, *data auditability* - data can be linked to company performance and profits, *data timeliness* - timeliness can be measured as the time between when information is expected and when it is readily available for use, *data orderliness* - measured by degree of data randomness and entropy, *data uniqueness*- a measure of unwanted duplication existing within or across systems for a particular field, record, or data set.

Data quality can deteriorate due to several factors such as sensor faults, bias, drift, freezing, and precision degradation [3] often due to aging and operating in harsh and uncertain environments. For instance, temperature variations introduce bias in force sensors, electromagnetic noise can affect accelerometer sensor readings, intermittent connectivity loss caused by physical barriers can introduce missing data over time, and unreliable communication protocols can cause intermittent connectivity loss. Additionally, sensor data that has been duplicated/transformed using unconventional methods and converted from analog to digital has discrepancies. Our faith in and dependence on CPS/IoT are diminished by poor data quality. For instance, raising a false alarm of a potential cardiac arrhythmia or a heart attack based on poor quality data from on-body ECG sensors is highly undesirable. Failing to stop a machining process leading to defects in products or tool wear and tear leads to tremendous amounts of waste in the manufacturing industry [4] which is estimated to be several hundred million tons per year worldwide.

How can software engineering and artificial intelligence (AI) help manage and tame data quality issues in CPS/IoT? This is the question we aim to investigate in this workshop SEA4DQ[1]. Emerging trends in software engineering need to take data quality management [5] seriously as CPS are increasingly data-centric in their approach to acquiring and processing, and sharing data [6], [7] along the edge-fog-cloud (EFG) continuum. The EFG continuum presents the challenge of data undergoing transformation from analog to digital and travelling through heterogeneous software and hardware spread across sensors, actuators, edge processing devices, local fog infrastructure, and global cloud infrastructure at, very often, sub-microsecond sampling frequencies. There is a need for novel software/hardware architectures for the EFG continuum to handle and process high-velocity multivariate sensor data with minimal data corruption owing to potentially harsh environmental conditions or noise sensors are exposed to, lack of adequate storage/computational resources at the edge, limited battery life, latency, security attacks [8], [9] and losses in connectivity between for instance the resource-constrained edge and the cloud. There is a need to manage the traceability of different versions of data produced and consumed by different components in a CPS minimizing inconsistencies along the EFG continuum. We need new approaches to define and execute test cases to verify data quality in CPS along the technologically diverse EFG continuum. These software engineering techniques need to interact hand in hand with AI models to detect anomalies in data quality within both short-term streaming data and long-term historical data and to repair erroneous data [10], replace missing data, and detect ethical issues such as bias in data from CPS. For instance, we may ask what is most optimal for a given CPS: deploying AI models for data quality in the resource-constrained edge or the resourceful cloud? In addition, we can also look at data quality in the social and distributed dimension. How can we ensure data quality between multiple CPS operating in distributed network? Can data quality metrics in CPS be recorded in distributed ledger technologies or a block chain to increase trust and reliability in data transferred across CPS? Can up-stream data users gain of knowing the quality of data describing the previous steps of the production line of the physical products, measurements in those steps, and what is the quality of the descriptions of the raw material entering the production line, etc.? Finally, validated approaches to manage data quality in CPS need to be used in certification of CPS and ideally contribute to standardization efforts [2].

The SEA4DQ workshop series originated from common research interests and international cooperation efforts, especially of Horizon 2020 EU projects InterQ[2] and the Dat4.ZERO[3] on data quality for Industry4.0. SEA4DQ'22 was a successful event organized on 17 November 2022, collocated with the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC / FSE) 2022 in Singapore. Compared to the first edition SEA4DQ 2021 [11], which featured one keynote, five presentations, and one panel discussion, the second edition SEA4DQ 2022 has evolved significantly with two keynotes, eight paper submissions, six presentations, and one panel discussion. Eight papers submitted to SEA4DQ 2022 had gone through a rigorous review process by the Program Committee, with three/four reviewers per paper. Submissions of PC members were treated with clear declaration of conflict of interest and decided by the PC chair without conflict of interest. In the end, based on the reviews, the PC had decided to accept two full papers, one work-in-progress paper, and two position papers to be included in the workshop proceedings [12]. Five paper presentations are part of the SEA4DQ 2022's program together with two keynotes, one project presentation (InterQ), and a panel discussion.

A summary of the keynotes and the talks are presented in Section 2. The workshop ended with an interesting panel discussion moderated by Beatriz Bretones Cassoli and Nicolas Jourdan that we summarize in Section 3. We conclude in Section 4 with the achievements of SEA4DQ 2022 and set up the goals for the next workshop SEA4DQ 2023.

2. SUMMARY OF KEYNOTE AND TALKS

This workshop hosted two keynotes, six presentations, and a panel discussion. The topics of interest for the workshop included:

- Software/hardware architectures and frameworks for data quality management in CPS.
- Software engineering and AI to detect anomalies in CPS data.
- Software engineering and AI to repair erroneous CPS data.
- Software tools for data quality management, testing, and profiling.
- Public sensor datasets from CPS (manufacturing, digital health, energy, etc.).
- Distributed ledger and blockchain technologies for quality tracking.
- Quantification of data quality hallmarks and uncertainty in data repair.
- Sensor data fusion techniques for improving data quality and prediction.
- Augmented data quality.
- Case studies that have evaluated an existing technique or tool on real systems, not only toy problems, to manage data quality in cyber-physical systems in different sectors.
- Certification and standardization of data quality in CPS/IoT.
- Approaches for secure and trusted data sharing, especially for data quality, management, and governance in CPS/IoT.
- Trade-offs between data quality and data security in CPS/IoT.

Most of these topics have been covered during the workshop. Phu Nguyen and Sagar Sen from SINTEF (Norway) opened the workshop that focuses on "*How can software engineering and artificial intelligence (AI) help manage and tame data quality issues in CPS/IoT?*". The opening address highlighted the importance of taming the data quality problem with software engineering and AI. We presented statistics from the workshop where we received eight submissions of which two full papers, two short papers, and one work-in-progress paper were accepted. The affiliations of authors of accepted papers were from Norway, Austria, Pakistan, and Singapore. The PC members who reviewed these articles were from Finland, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden, United Kingdom, and USA. We hope to increase the national and gender diversity by reaching out to PC members in North/South America, Africa, and Asia for the PC of the next edition of the SEA4DQ.

The first keynote was delivered by Prof. Andreas Metzger from the department of Software Systems Engineering (SSE) at Universität Duisburg-Essen. His talk entitled "Online Reinforcement Learning for Self-adaptive Systems" presented *data quality issues* encountered during reinforcement learning in self-adaptive systems. These data quality included *sparsity, drift,* and *opaqueness*. Sparsity refers to the large size of discrete adaptation space of which very few are optimal adaptations. The sparsity problem is addressed by specifying a constrained space of adaptations using feature models. Metzger showed

[1] https://sea4dq.github.io/

[2] https://interq-project.eu/

[3] https://dat4zero.eu/

that feature models guided reinforcement learning constrained the adaptation domain and converged faster than random exploration. Data drift is the second challenge he addressed where non-stationarity of the environment renders adaptation policies in reinforcement learning sub-optimal over time. The third data quality issue they deal with is data opaqueness where it is very hard to understand decision-making based on data points for reward, state, and action in reinforcement learning. Metzger presented explainable AI for reinforcement learning where they visualize the influence of rewards of sub-agents on the composed decision.

The keynote set the stage for the morning session at SEA4DQ. The first talk was by Xiang Ma on "Data Quality Issues in Solar Panels Installations". Xiang presented SINTEF's rooftop installation of solar panels from different manufacturers and a data pipeline to acquire integrate data from the solar panels, AC/DC inverters, local weather stations and the weather forecast. The authors (Maryna Waszak, Terje Moen, Sølve Eidnes, Alexander Stasik, Anders Hansen, Gregory Bouquet, Antoine Pultier, Xiang Ma, Idar Tørlen, Bjørn Rune Henriksen, Arianeh Aamodt, Dumitru Roman) address data quality issues of missing data and inconsistent timing through average and linear interpolation. Furthermore, they use fault detection algorithms to identify unknown conditions to discard data for instance when solar panels are moved around. At this stage repairing data is automatically hard as this would require acquisition for many years. The next talk was by Jørgen Stang on "Data Quality as a Microservice - an ontology and rule-based approach for quality assurance of sensor data in manufacturing machines". Jørgen from DNV's data quality group (with Dirk Walther, Per Myrseth) presented the core idea of representing data quality requirements in the form of an ontology. These requirements were represented as expectations on generic signals and more specific signals that inherit all the rules from generic signals. Given a dataset and set of variables ontological reasoning is important to understand what data quality requirements apply. The third talk in the morning session was by Valentina Golendukhina on "Preliminary Findings on the Occurrence and Causes of Data Smells in a Real-World Business Travel Data Processing Pipeline", co-authored by Harald Foidl, Michael Felderer and Rudolf Ramler. Valentina presented encoding, consistency, and believability smells along with causes for them and methods to detect such smells. Andreas Metzger from the audience pointed out that the concept of data smells needs to be clarified with respect to the concept of code smells used in the community. The last article in the morning session was by Mohammed Azmi Umer (with Aditya Mathur and Muhammad Taha Jilani) on "Effect of Time Patterns in Mining Process Invariants for Industrial Control Systems: An Experimental Study". Mohammed presented the idea of invariant mining based on frequent itemset mining to extract data quality invariants.

The second keynote after lunch was delivered by Fouste Khomh, who is a Professor of Software Engineering at Polytechnique Montréal (Canada) where he leads the SoftWare Analytics and Technologies (SWAT) Lab. Foutse's keynote address was on "Data Quality and Model Under-Specification Issues". Foutse presented meta-heuristic search and metamorphic operators to augment a dataset to training machine learning models to mitigate the problem of model under-specification due to limited data. Foutse highlighted the need for meaningful metamorphic relations for multimodal data to obtain better datasets. The second keynote elevated the energy level for the afternoon session where the first talk on "Data Quality Issues for Vibration Sensors: A Case Study in Ferrosilicon Production" was again delivered by Xiang Ma on behalf of his colleagues (Dumitru Roman, Antoine Pultier, Ahmet Soylu, Alexander G.Ulyashin). Xiang presented the challenges of acquiring vibration data in harsh production environments.

The last talk in the afternoon session was by Beatriz Cassoli and Nicolas Jourdan on the sponsoring project InterQ. They presented an overview of the InterQ project on zero-defect manufacturing and presented the learning factory at TU Darmstadt where data is acquired from a production line to produce piston rods. The learning factory provides a comprehensive platform to run controlled experiments in manufacturing and evaluate the inter-relationship between process and product data and the quality of the data itself.

The workshop ended with fruitful discussions in the afternoon panel, summarized in the following section.

3. SUMMARY OF DISCUSSIONS

The panel discussion in SEA4DQ was co-moderated by Beatriz Bretones Cassoli and Nicolas Jourdan from PTW, TU Darmstadt, Germany. The panelists were:

- Prof. Andreas Metzger from the University of Duisburg-Essen, Germany
- Prof. Foutse Komhh from Polytechnique Montréal, Canada
- Dr. Sallam Abualhaija from the University of Luxembourg
- Dr. Sagar Sen from SINTEF, Norway

The panelists, together with the workshop participants, actively discussed the current challenges and the visions for the future of data quality and reliable AI applications based on thought-provoking questions suggested by the moderators Beatriz Bretones Cassoli and Nicolas Jourdan. The questions and how they were addressed are summarized below:

How do you model and validate data quality requirements in your domain?

Sallam started answering the first question by explaining data quality concerns in the domain of text document processing for automated quality assurance and regulatory compliance checks. The processed documents typically include requirement specifications or legal documents. Data labelling is done mostly manually, and quality assurance is done through inter-rater agreements. Nicolas remarked that inter-rater agreement is also often used in computer vision applications that are common in manufacturing. Sagar answered the question by providing his perspective on the medical and manufacturing domains. In the medical domain, data quality is crucial as it is used in applications that directly influence the patient's diagnosis and treatment. The requirements mostly come from current standards and norms. In manufacturing, data requirements are derived from application-specific needs such as data latency, formatting standards and data quantity.

What aspects of data quality can be addressed purely by automation using SE and AI in your domain?

In response to this question, Sallam highlighted that, while some structural inconsistencies may be detected purely automatically, a lot of manual effort is still spent on cleaning the dataset afterward. Especially issues such as ambiguity in textual requirements may be resolved automatically while issues like missing or incomplete data still require a human's attention. Sagar mentioned research results from the InterQ project which, among other topics, investigated the automatic identification of data quality issues in manufacturing datasets and data streams. While SE and AI were successfully employed for the detection of data quality issues, the resolution of these issues was found to require human attention in most cases. The panelists thus formed a consensus regarding the use of automation for solving data quality issues in datasets: While SE and AI can help to identify data quality issues and partially resolve them, human attention is still required for a significant part of the possible data quality problems. This highlights the importance of educating and working with practitioners who collect the data and use the AI applications which is discussed in detail in the last question of the panel discussion.

In your experience, how does data quality influence the reliability of AI applications?

The panelists answered this question by agreeing that AI/machine learning solutions' performance and reliability strongly depend on the quality of the data used for developing and running the application. Andreas noted that the data preparation for training machine learning models usually corresponds to 80% of the project time and is decisive for performance results. Beatriz remarked that depending on the quality and characteristics of the dataset at hand, the machine learning task must be reframed, and the problem tackled from a different perspective, such as in the case of highly imbalanced datasets, and the consideration between classification and anomaly detection approaches. Foutse drew attention to the fact that data quality is paramount when using small datasets. Andreas, Foutse, Sallam and Sagar further discussed the role of the domain experts. They recognized the need for domain knowledge to define the AI task and evaluate the quality of the data used in the solution.

How can we make AI models robust to data quality issues?

As this question is highly related to the keynote presentations by Foutse and Andreas, the discussion mainly revolved around the concepts mentioned there. Three main takeaways for building robust and reliable AI models were identified in this part of the session: First, it is important to tackle the problem of insufficient data quality and quantity proactively in the very early stages of AI projects. This involves identifying requirements and working with practitioners to set up appropriate data gathering or measurement methods and thus reducing data quality issues during the dataset acquisition phase. Practitioners and users must be educated about the influence of data quality on the overall reliability and performance of AI models to enable efficient development and safe usage thereof. Second, methods in the scope of robustness testing using, e.g., data augmentation strategies may be used to identify weak spots in existing models and to generate targeted additional training data to increase the reliability of these models. Lastly, AI applications should be aware of their limitations regarding input data distributions and data quality issues and thus warn users accordingly if they are used outside of their designed operating conditions. Sagar suggested the use of uncertainty estimation methods to assess the confidence of AI models in their predictions which can be used to reason about their reliability in each scenario.

4. CONCLUSION & WORKSHOP'S FUTURE

SEA4DQ 2022 has again provided researchers and practitioners a forum for exchanging ideas, experiences, understanding of the problems, visions for the future, and promising solutions to the problems of data quality in CPS/IoT. Compared to SEA4DQ 2021, the second edition SEA4DQ 2022 has evolved significantly with two keynotes, eight paper submissions, six presentations, and one panel discussion. Two full papers, one work-in-progress paper, and two position papers have been included in the workshop proceedings.

Although SE and AI can aid in identifying and partially resolving data quality concerns, a sizable portion of such issues still require human attention. There was clear consensus that the effectiveness and dependability of AI/machine learning solutions are highly influenced by the caliber of the data utilized to create and maintain the application. Applications using AI should be cognizant of their limits with relation to input data distributions and data quality concerns. The workshop ended with fruitful discussions in the afternoon panel where the panelists actively discussed the current challenges and the visions for the future of data quality and reliable AI applications.

For the next edition of the workshop, we will try to draw in additional practitioners, researchers, and important figures in the CPS, IIoT, and Industry 4.0 spheres. We hope to increase the national and gender diversity by reaching out to PC members in North/South America, Africa, and Asia for the PC of the next edition of the SEA4DQ. Finally, we will encourage participants to submit extended versions of their work for a special issue in a suitable journal.

Let us look forward to SEA4DQ 2023!

5. ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 958357 (InterQ), and Grant Agreement No. 958363 (DAT4.Zero).

We would like to thank the program committee members and all reviewers for their work in evaluating the submissions. We also thank the ESEC / FSE 2022 organizers, student volunteers for their assistance in the preparation and operation of the workshop and the editors of ESEC / FSE 2022 for help in publishing the workshop proceedings.

6. REFERENCES

[1] D. International, *DAMA-DMBOK: data management body of knowledge*. Technics Publications, LLC, 2017.

[2] DNV, *Recommended Practice 0497 - Data quality assessment framework*. DNV, 2017.

[3] S. U. Jan, Y. D. Lee, and I. S. Koo, 'A distributed sensor-fault detection and diagnosis framework using machine learning', *Information Sciences*, vol. 547, pp. 777–796, Feb. 2021, doi: 10.1016/j.ins.2020.08.068.

[4] D. Powell, M. C. Magnanini, M. Colledani, and O. Myklebust, 'Advancing zero defect manufacturing: A state-of-the-art perspective and future research directions', *Computers in Industry*, vol. 136, p. 103596, Apr. 2022, doi: 10.1016/j.compind.2021.103596.

[5] B. B. Cassoli, N. Jourdan, P. H. Nguyen, S. Sen, E. Garcia-Ceja, and J. Metternich, 'Frameworks for data-driven quality management in cyber-physical systems for manufacturing: A systematic review', *Procedia CIRP*, vol. 112, pp. 567–572, Jan. 2022, doi: 10.1016/j.procir.2022.09.062.

[6] H.-H. Nguyen, P. H. Phung, P. H. Nguyen, and H.-L. Truong, 'Context-driven Policies Enforcement for Edge-based IoT Data Sharing-as-a-Service', in *2022 IEEE International Conference on Services Computing (SCC)*, 2022, pp. 221–230. doi: 10.1109/SCC55611.2022.00041.

[7] S. Alshehri, O. Bamasaq, D. Alghazzawi, and A. Jamjoom, 'Dynamic Secure Access Control and Data Sharing Through Trusted Delegation and Revocation in a Blockchain-Enabled Cloud-IoT Environment', *IEEE Internet of Things Journal*, pp. 1–1, 2022, doi: 10.1109/JIOT.2022.3217087.

[8] G. Erdogan, E. Garcia-Ceja, Å. Hugo, P. H. Nguyen, and S. Sen, 'A Systematic Mapping Study on Approaches for AI-Supported Security Risk Assessment', in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 755–760. doi: 10.1109/COMPSAC51774.2021.00107.

[9] T. Rajmohan, P. H. Nguyen, and N. Ferry, 'A decade of research on patterns and architectures for IoT security', *Cybersecurity*, vol. 5, no. 1, p. 2, Jan. 2022, doi: 10.1186/s42400-021-00104-7.

[10] S. Sen, E. J. Husom, A. Goknil, S. Tverdal, P. Nguyen, and I. Mancisidor, 'Taming Data Quality in AI-Enabled Industrial Internet of Things', *IEEE Software*, vol. 39, no. 6, pp. 35–42, 2022, doi: 10.1109/MS.2022.3193975.

[11] P. H. Nguyen *et al.*, 'Software Engineering and AI for Data Quality in Cyber- Physical Systems - SEA4DQ'21 Workshop Report', *SIGSOFT Softw. Eng. Notes*, vol. 47, no. 1, pp. 26–29, Jan. 2022, doi: 10.1145/3502771.3502781.

[12] P. Nguyen, S. Sen, and M. C. Magnanini, *Proceedings of the 2nd International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things*. CM. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3549037