# Software Engineering and AI for Data Quality in Cyber-Physical Systems – SEA4DQ’21 Workshop Report

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

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

Per Myrseth
DNV AS
Oslo, Norway
per.myrseth@dnv.com

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

Odd Myklebust
SINTEF Manufacturing
Trondheim, Norway
odd.myklebust@sintef.no

DOI: 10.1145/3502771.3502781
https://doi.org/10.1145/3502771.3502781

## ABSTRACT

Cyber-physical systems (CPS) have been developed in many industrial sectors and application domains in which the quality requirements of data acquired are a common factor. Data quality in CPS can deteriorate because of several factors such as sensor faults and failures due to operating in harsh and uncertain environments. How can software engineering and artificial intelligence (AI) help manage and tame data quality issues in CPS? This is the question we aimed to investigate in the SEA4DQ workshop. Emerging trends in software engineering need to take data quality management seriously as CPS are increasingly data-centric in their approach to acquiring and processing data along the edge-fog-cloud continuum. This workshop provided 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. Examples of topics include software/hardware architectures and frameworks for data quality management in CPS; software engineering and AI to detect anomalies in CPS data or to repair erroneous CPS data. SEA4DQ 2021, which took place on August 24th, 2021 was a satellite event of the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC / FSE) 2021. The workshop attracted 35 international participants and was exciting with a great keynote, six excellent presentations, and concluded on a high note with a panel discussion. SEA4DQ was motivated by the common research interests from the EU projects for Zero-Defects Manufacturing such as InterQ and Dat4.Zero.

## 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) have been developed 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. We can list several examples: CPS in manufacturing have been used to acquire high frequency data from machine tools such as Computer Numerical Control (CNC) machines for predictive maintenance yearning towards zero defects manufacturing. CPS in digital health have been used to acquire data from body sensors such as in overnight polysomnography to predict obstructive sleep apnea estimated to affect over a billion people worldwide. CPS in the automotive industry has been used to acquire data from semi-autonomous and connected vehicles for safe navigation and traffic management. CPS in the energy sector have been used for acquiring data about energy production and consumption for smart metering and lowering our household carbon footprint.

The quality of the data acquired and used for decision support is a common factor across industrial sectors that underpins our *reliability on* and *trust of* CPS. There exist many classifications for data quality in the literature [1]. 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 [2] often due to aging and operating in harsh and uncertain environments. For instance, electromagnetic noise can affect accelerometer sensor readings, temperature variations introduce bias in force sensors, intermittent loss in connectivity due to physical barriers or unreliable communication protocols introduce missing data over periods of time. Moreover, there are inconsistencies in sensor data that are transformed from analog to digital and duplicated/transformed by non-standard approaches. Poor data quality reduces our trust and reliance on CPS. 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 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? This is the question we aim to investigate in this workshop SEA4DQ. Emerging trends in software engineering need to take data quality management seriously as CPS are increasingly data-centric in their approach to acquiring and processing data 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 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, 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 I 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.

The SEA4DQ 2021 workshop originated from common research interests and international cooperation efforts, especially of Horizon 2020 EU projects InterQ[1] and the Dat4.ZERO[2] on data quality for Industry4.0. It was a successful event organized on 24 August 2021, attracting 35 participants. The workshop has provided researchers and practitioners with a forum for exchanging ideas, experiences, understanding of the problems, visions for the future, and promising solutions to the problems in data quality in Cyber-Physical Systems, especially for Zero-Defect Manufacturing. A summary of the keynotes and the talks are presented in Section 2. The workshop ended with an interesting panel discussion catalyzed by Sagar Sen and Frank Westad that we summarize in Section 3. We conclude in Section 4 with goals for the next workshop.

## 2. SUMMARY OF KEYNOTE AND TALKS

This workshop hosted one keynote and six presentations. 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
- Approaches for secure and trusted data sharing, especially for data quality, management, and governance in CPS
- Trade-offs between data quality and data security in CPS

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?*". In his keynote, Per Myrseth from DNV (Norway) emphasized the importance of "*Data Quality Focus as a Competitive Advantage",* based on his great experience for Data Management, Data Science and Assurance of Digital Assets [3]. Indeed, the recommended practice by DNV for Data quality assessment framework was mentioned in the following talk about the "*Development of Data Quality Management System for Ship IoT Data – Perspective of Ship Owner and Operator*" by Putu Hangga and Shogo Yamada from the Maritime Technology Group, NYK Remote Diagnostic Center, Japan.

Next, Frank Westad and Torbjørn Pedersen from Idletechs (Norway) shared their experience for "*Representative Sampling*", which is an important basis for acquiring high quality data. Then, Phu Nguyen, Arda Goknil, Karl John Pedersen, and Dimitra Politaki represented a group of the authors from the InterQ project to present their systematic assessment on the state of the art of data quality. Their presentation entitled "*the Preliminary Results of A Systematic Review of Data Quality for CPS, IoT, or Industry 4.0 Applications*". The morning session was concluded with a talk by Dimitra Politaki from Inlecom Systems (Greece) on "*how to combine AI models for Anomaly Detection in Manufacturing Time Series Data*".

The afternoon session started with an inspiring talk entitled "*Injection molding supervision and plastic part quality assurance*" by Ronan Le Goff from IPC – Centre Technique Industriel de la Plasturgie et des Composites (France). Ronan presented how their data science approach has allowed replacing an expensive sensor system with a different and much less expensive one that can do the same job for quality assessment in their manufacturing process. Last but not least, Sonia Jimenez from the International Data Spaces Association (IDSA, Germany) presented "*how the International Data Spaces´ approach for secure and trusted data sharing contributes to ensuring data quality*". This talk shows a great vision of data sharing with high quality and how it has been realized with the IDSA's reference architecture model.

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

[2] https://dat4zero.eu/

## 3. SUMMARY OF DISCUSSIONS

Panel discussions in SEA4DQ was co-moderated by Frank Westad from IDLETECHS[3], Norway and Sagar Sen from SINTEF Digital[4], Norway. The panelists were Per Myrseth from DNV[5], Norway, Dimitra Politaki, INLECOM[6], Greece, and Sonia Jimenez from the International Data Spaces Association[7]. All the panelists and the participants actively discussed the current challenges and the visions for the future of data quality based on thought provoking questions suggested by the moderators Sagar Sen and Frank Westad. The questions and how they were addressed are summarized below:

*Are there off-the-shelf solutions for automatic data alignment for sensor-fusion and multi-step models, or is there always a need for humans-in-the-loop?* This question from Frank stems from his experience with batch alignment of data and his scepticism towards the use of interpolation in sensor data. Dimitra believes that complete automation of data alignment is still in its in-fancy and will require humans in the loop specially to prepare data for AI modelling and deep learning. Per has seen several tools that automate data cleaning for customer databases but is unaware of similar tools for sensor data. He also mentioned the need to synchronize *tags and identifiers* used with digital twins in maritime, electricity, oil & gas such as drawings, videos, radar analytics, and a lot of sensor data. Sagar suggested that sending synchronization signals to all sensors with a unique representation can be used to align data. He added that tools like Adobe Premiere automate the alignment of video and audio data when the video signal contains a weaker version of the audio signal. He believes that looking for clusters of data characterizing similar events can help automate data alignment across time series. However, the problems with missing data and varying sampling frequencies need to be dealt with to make sensor data comparable.

*How should one analyse and monitor data with a mix of controlled and observed process parameters, especially for time-dependent (batch) processes?* Frank gave context to the question where process targets in temperature say 37 degree C or pressure 2 millibar also experience random fluctuations for the same parameters. Changing parameters on the other hand affects the process. Should one change parameters or keep them stable. Ronan from the audience said that anomaly detection as presented earlier by Dimitra can be useful in figuring out when the data is good enough for analysis. Dimitra suggested that time series data should be looked at based on their properties such as stationarity. Sagar added that in manufacturing for instance periodicity of production cycles can help split batch data into cycles and their statistical and topological properties may then be analysed and compared. Frank followed by suggesting that one could impose seasonal autoregressive integrated moving average (SARIMA) models on the data.

*To what extent are security constraints such as firewalls a hindrance for deploying on-line solutions and data quality (re. IT/OT infrastructure, cyber security)?* Phu provided context to this question by exemplifying the conflict between security and data quality requirements. For instance, encryption can introduce data latency and hence affect data currency. Privacy constraints can prevent the use of some personal data. Per concurred that the flow of data is often blocked or slowed down by firewalls. Putu explained how they deal with data with multiple owners. They use virtual private cloud (VPC) and content delivery networks such as Azure Frontdoor along with their applications. Phu also mentioned that Blockchain databases can improve diagnosis of data quality faults and root cause analysis by allowing traceability during the sharing process.

*What is the state-of-art for deploying parallel models for assessing the consensus from prediction with various ML/AI methods - and is this a relevant topic (re. self-adaptive systems)?* Ronan agreed that the question was relevant as manufacturing systems require different forms of intelligence in parallel. For instance, an AI that can predict tool wear or good or bad parts would need a parallel model to explain why production failed. Frank mentioned how explainable AI can reveal AI models fail by distinguishing between wolves and huskies based on the snow in the background. Sagar elaborated on the evaluation of explainable AI models based on methods such as simulatability and counterfactual simulation where humans learn to predict the behaviour of the model (TP, TN, FP, FN) both for its positive aspects and its flaws. Phillip and Sagar added a little more to the discussion of evaluation AI models using fake images using generative adversarial networks and adversarial testing.

*Does the data minimization principle in GDPR apply to data quality in CPS? If it does how should one manage the data to that end? If not, why?* Sagar presented the context where we use data for an intended purpose for a given amount of time and then destroy it as prescribed by the data minimization principle. Dimitra added that anonymization of personal data is an approach to handle GDPR. She did not see why we should destroy or stop using data if it is not sensitive. Sagar explained that copies of personal health data are often supposed to be deleted after its intended use and requested again. However, in the case of non-personal data such as in manufacturing, data when used for an intended purpose such as predictive maintenance can also be used to malign a company. Per shared his experience where DNV uses data to show how ships are in good condition but are selective about general data use to protect a company’s reputation.

*What should one do with dark data acquired through CPS? How can we maintain/improve the quality of dark data so that it stays relevant?* Dimitra insisted on refining the use case for data collection, identifying low-quality data, and deleting it automatically. However, Sagar said that companies typically store the data first before figuring out what to do with it resulting in mountains of dark data. Per added that this trend is increasingly common to storage becoming extremely cheap. It is far most expensive to clean and choose data. Furthermore, he emphasized that we don’t think about metadata anymore which has triggered unstructured big data storage. Sonia took the middle ground where she said that although it is easier to buy more capacity, we need to put the right processes in place to handle the data in addition to determining the intended purpose for the data. Putu added that it is true and a very confusing problem. For him, the definition of dark data is very vague. For instance, email or log data is dark data but what should one do with it. They keep data because it's part of a regulation. For example, ship clinical records from the last 15 years are needed by insurance companies. Their current movement is to digitize such data and store it, but they are unsure about how to deal with it. Per shared that key data of a physical asset often will live much longer than the asset itself. Some data needs to be stored for 100 years. We need a retention policy and that’s part of data management. It is a hard decision to archive data due to the costs, but some data may be worth gold in the future. We need to take that risk. Erik from the audience shared his experience with leaving a low personal digital footprint by deleting most emails. He still

[3] https://idletechs.com/

[4] https://www.sintef.no/

[5] https://www.dnv.com/services/data-management-and-quality-144663

[6] https://inlecom.eu/

[7] https://internationaldataspaces.org/

receives advertisements, but they are less and less targeted. One may also argue that its perhaps better to obtain targeted ads than ads that are irrelevant.

*Which way to go – federated learning on local data and exchange of parameters or international data spaces?* Sagar elaborated on the question: should we aim to build global AI models with model parameter sharing instead of data sharing or focus on data sharing as prescribed by IDSA? Sonia emphasized that we need to ensure trust while sharing data and we need to share more data. There must be a focus on balancing both the legal and technical dimensions. Sagar asks what happens when a far-right government comes into power and wants to use the data for a completely different purpose. Sonia emphasized that a threshold of trust needs to be established through a negotiation process. Sagar asked what if some data is critical and requires urgency of use. Sonia mentions that all IDSA members are trusted partners who sign a contract and it only becomes complex when the data contains personal information due to GDPR. Per concluded by stating that the principle of IDSA is good but in a society based on capitalism doesn't work that way. Some actors will not accept fair play, and some will take positions to earn money where data is the raw material where they will make money on. We need to be careful about where we are all friends.

## 4. CONCLUSION & WORKSHOP'S FUTURE

SEA4DQ 2021 provided 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. The workshop also provided a platform for researchers and developers of tools for data quality to work together to identify the problems in the theory and practice of data quality in CPS and to set an agenda and lay the foundation for future development.

The problem of data synchronization and alignment or simply put pre-processing is the elephant in the room that software engineering and AI approaches should aim to automate. There was clear consensus that this activity requires human intervention and very little literature deals with finding software engineering and AI patterns to address the problem.

The interplay between security and data quality has been a pervasive topic in SEA4DQ. Blockchain databases, federated learning, smart contracts for data sharing were all brought up to be highly relevant topics. On the other hand, evaluating the explainability of AI models is one key area of research to increase trust in AI-based software that makes predictions based on big data from CPS. There was particular interest in dealing with *dark data* and data minimization, which can be an interesting avenue for software engineering and AI researchers.

We will try to attract more practitioners, academics, and significant players in the CPS, IIoT, Industry 4.0 spaces for the next edition of the workshop. Another goal is to encourage collaboration among relevant European research projects and integrate individuals who want to participate in the program committee.

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 2022!

## 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).

## 6. REFERENCES

[1] International, D., *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*. 2017: Technics Publications, LLC.

[2] Jan, S.U., Y.D. Lee, and I.S. Koo, *A distributed sensor-fault detection and diagnosis framework using machine learning.* Information Sciences, 2021. **547**: p. 777-796.

[3] DNV, *DNV GL-Recommended Practice 0497 - Data quality assessment framework*, in *DNV GL-Recommended Practice*, DNV, Editor. 2017, DNV.