2nd International Workshop on Data Quality Assessment for Machine Learning

Hima Patel*
IBM Research, India
himapatel@in.ibm.com

Fuyuki Ishikawa
NII, Japan
f-ishikawa@nii.ac.jp

Laure Berti-Equille
IRD, France
laure.berti@ird.fr

Nitin Gupta
IBM Research, India
ngupta47@in.ibm.com

Sameep Mehta
IBM Research, India
sameepmehta@in.ibm.com

Satoshi Masuda
IBM Research, Japan
smasuda@jp.ibm.com

Shashank Mujumdar
IBM Research, India
shamujum@in.ibm.com

Shazia Afzal
IBM Research, India
shaafzal@in.ibm.com

Srikanta Bedathur
IIT Delhi, India
srikanta@cse.iitd.ac.in

Yasuharu Nishi
UEC, Japan
yasuharu.nishi@uec.ac.jp

ABSTRACT

The 2nd International Workshop on Data Quality Assessment for Machine Learning (DQAML'21) is organized in conjunction with the Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). This workshop aims to serve as a forum for the presentation of research related to data quality assessment and remediation in AI/ML pipeline. Data quality is a critical issue in the data preparation phase and involves numerous challenging problems related to detection, remediation, visualization and evaluation of data issues. The workshop aims to provide a platform to researchers and practitioners to discuss such challenges across different modalities of data like structured, time series, text and graphical. The aim is to attract perspectives from both industrial and academic circles.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

Data Quality, Data Assessment, Machine Learning

ACM Reference Format:
Hima Patel, Fuyuki Ishikawa, Laure Berti-Equille, Nitin Gupta, Sameep Mehta, Satoshi Masuda, Shashank Mujumdar, Shazia Afzal, Srikanta Bedathur, and Yasuharu Nishi. 2021. 2nd International Workshop on Data Quality Assessment for Machine Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3469468

*Author names are arranged alphabetically.

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3469468

1 BACKGROUND

In the past decade, AI/ML technologies have become pervasive in academia and industry, finding their utility in newer and challenging applications. While most efforts are focused towards building better, smarter and automated ML models, comparatively lesser attention has been paid to systematically understand the challenges in training data and assess its quality before it is fed to an ML pipeline. Issues such as incorrect labels, synonymous categories in a categorical variable, heterogeneity in columns, class imbalance, etc. that might go undetected by standard data pre-processing modules can lead to increased model complexity or cause sub-optimal model performance. So data preparation [1–5] is being rightly been called out as one of the most time-consuming and challenging step in an AI lifecyle. This is because the quality of data remains unknown at the acquisition stage triggering an iterative debugging process mostly leveraging the experience of a data scientist. Data quality assessment for ML is therefore an important step to estimate and improve the data quality by identifying anomalies and applying corresponding remediation algorithms. Interesting issues that spawn this thread are related to optimizing humans-in-the-loop, offering adequate explanations, supporting smart visualizations, and actionable insights amongst other topics listed in Section 1.1.

The goal of this workshop is to attract researchers working in related fields of data acquisition, data labeling, data quality, data preparation and AutoML areas to understand how the data issues, their detection and remediation will help towards building better models. This workshop invites researchers from academia and industry to submit novel propositions for systematically identifying and mitigating data issues for making data AI ready.

1.1 Topics of Interest

Methods of data assessment can change depending on the modality of the data. In this workshop we focus on data quality assessment for different modalities: structured (or tabular) data, unstructured data (such as text, logs, images), graph structured data (relational, network), time series data, spatio-temporal data etc. Following is the list of topics that are of interest to this workshop:

- Algorithms for assessment of data quality issues relevant to ML
- Automatic remediation of data quality issues
- Human-assisted data cleaning and remediation
- Automated data cleaning workflows
- Explainability and interpretability of quality assessment
- Interactive debugging of data
- Smarter data visualizations for high dimensional data
- Evaluation techniques for data quality assessment
- Real world use cases and applications of data quality assessment
- Novel interfaces to assist human-in-the-loop intervention
- Representative sampling for high dimensional data
- Detection of bias and personal information in data
- Label noise detection, explanation and incorporating feedback
- Auto ordering of datasets on difficulty level with explanations
- Handling corrupted, missing and uncertain data
- Outlier (or anomaly) detection and mitigation in data
- Addressing Class Imbalance in data
- Syntactic Data Validations

2 WORKSHOP PROGRAM

The workshop is organised as a half day meeting (~ 3.5 hours). All details of the keynote/invite talks, accepted papers and the overall program details can be found on the workshop website: **http://data-readiness-kdd-2021.mybluemix.net/**.

3 PROGRAM COMMITTEE

We appreciate the reviewers' efforts and would like to thank the members of the program committee for their valuable support: Shanmukh Chaitanya (IBM Research), Ruhi Sharma Mittal (IBM Research), Pranay Lohia (IBM Research), Aniya Agarwal (IBM Research), Lokesh N. (IIT Bombay), Vitobha Munigala (IBM Research), Abhinav Jain (Videoken), Hideto Ogawa (Hitachi Ltd.), Fei Chang (McMaster University), George Papadakis (National Technical University of Athens), Evelyn Duesterwald (IBM), Shaikh Quader (IBM) and Heiko Mueller (New York University).

4 ORGANIZERS

- **Hima Patel** is a research manager at IBM Research AI India and leads a team of researchers that work in the area of making data ready for an AI lifecycle.

- **Fuyuki Ishikawa** is Associate Professor at Information Systems Architecture Science Research Division, and also Deputy Director at GRACE Center, in National Institute of Informatics, Japan.

- **Laure Berti-Equille** is a Research Director at IRD and leads a team of researchers. Her work is at the intersection of data management and machine learning with a focus on data quality, data cleaning and data preparation with more than 100 publications in major conferences and journals and three monographs.

- **Nitin Gupta** is a Research Scientist in Data and AI team at IBM Research AI Labs, India. He is currently leading technical efforts on developing "Data Quality for ML" framework to improve the overall performance of Machine Learning algorithms.

- **Sameep Mehta** heads Data and AI Research at IBM Research India. His focus is to develop novel algorithms and systems to help clients in their AI journey.

- **Satoshi Masuda** is a senior research staff at IBM Research Tokyo. His research interest is quality for AI, and his works includes testing question answering robot and verifying self driving car.

- **Shashank Mujumdar** is a Researcher in the Data and AI team working at IBM Research Lab, India. Currently, he's leading the work to systematically assess the quality of unstructured text data and log data.

- **Shazia Afzal** is a Research Scientist with the Data and AI team at IBM Research AI Labs, India. She is currently exploring the role of humans in the data quality and assessment pipeline for AI with a focus on explainability and transparency.

- **Srikanta Bedathur** is an Associate Professor and the DS Chair of Artificial Intelligence at the Department of Computer Science and Engineering at IIT Delhi. Until recently he worked as a research scientist at IBM Research as part of their AI Research team.

- **Yasuharu Nishi** is an Assistant Professor in the Department of Informatics at the University of Electro-Communications, Tokyo, Japan. He does research and consults for industries on software testing, quality, safety, process improvement, and TQM.

REFERENCES

[1] Bortik Bandyopadhyay, Sambaran Bandyopadhyay, Srikanta Bedathur, Nitin Gupta, Sameep Mehta, Shashank Mujumdar, Srinivasan Parthasarathy, and Hima Patel. 2021. 1st International Workshop on Data Assessment and Readiness for AI.. In *PAKDD (Workshops)*.
[2] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and Importance of Data Quality for Machine Learning Tasks. In *KDD*.
[3] Hima Patel, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, Vitobha Munigala, Naveen Panwar, Sambaran Bandyopadhyay, and Satoshi Musda. 2021. Data Quality for Machine Learning Tasks. In *KDD*.
[4] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv* (2020).
[5] Jinsung Yoon, Sercan Arik, and Tomas Pfister. 2020. Data valuation using reinforcement learning. In *ICML*.