TU Delft

Delft University of Technology

Workshop on Human-in-the-loop Data Curation

Demartini, Gianluca; Yang, Jie; Sadiq, Shazia

DOI
10.1145/3511808.3557498

Publication date
2022

Document Version
Final published version

Published in
CIKM 2022 - Proceedings of the 31st ACM International Conference on Information and Knowledge Management

Citation (APA)
Demartini, G., Yang, J., & Sadiq, S. (2022). Workshop on Human-in-the-loop Data Curation. In *CIKM 2022 - Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (pp. 5161-5162). (International Conference on Information and Knowledge Management, Proceedings). Association for Computing Machinery (ACM). https://doi.org/10.1145/3511808.3557498

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Workshop on Human-in-the-loop Data Curation

Gianluca Demartini
g.demartini@uq.edu.au
The University of Queensland
Brisbane, Australia

Jie Yang
j.yang-3@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Shazia Sadiq
s.sadiq@uq.edu.au
The University of Queensland
Brisbane, Australia

ABSTRACT

Although data quality is a long-standing and enduring problem, it has recently received a resurgence of attention due to the fast proliferation of data analytics, machine learning, and decision-support applications built upon the wide-scale availability and accessibility of (big) data. The success of such applications heavily relies on not only the quantity, but also the quality of data. Data curation, which may include annotation, cleaning, transformation, integration, etc., is a critical step to provide adequate assurances on the quality of analytics and machine learning results. Such data preparation activities are recognised as time and resource intensive for data scientists as data often comes with a number of challenges that need to be tackled before it can be used in practice. Data re-purposing and the resulting distance between design and use intentions of the data, is a fundamental issue behind many of these challenges. These challenges include a variety of data issues such as noise and outliers, incompleteness, representativeness or biases, heterogeneity of format or semantics, etc. Mishandling these challenges can lead to negative and sometimes damaging effects, especially in critical domains like healthcare, transport, and finance. An observable distinct feature of data quality in these contexts is the increasingly important role played by humans, being often the source of data generation and the active players in data curation. This workshop will provide an opportunity to explore the interdisciplinary overlap between manual, automated, and hybrid human-machine methods of data curation.

ACM Reference Format:
Gianluca Demartini, Jie Yang, and Shazia Sadiq. 2022. Workshop on Human-in-the-loop Data Curation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3511808.3557498

1 OBJECTIVES, GOALS, AND OUTCOMES

The need for new research effort on involving humans in the loop of the data curation process is exacerbated by the importance of developing methods that can scale to large amounts of data while also maintaining a human touch. This means designing processes that can deliver high level of transparency in the data curation process (e.g., explaining why certain values have been dropped), deal with ethical data challenges like the decision to use or discard certain attributes (e.g., applicants' gender) in decision making processes, and, overall, increase the quality and trust in the outcome. Given the cross-disciplinary nature of the problem, we will seek contributions from researchers in a variety of areas including databases, human-computer interaction, human computation and crowdsourcing, machine learning, and the broad area of AI as related to data curation, thereby creating a unique opportunity for researchers from these areas to bring together diverse perspectives and deliberate on the underlying challenges.

2 TARGET AUDIENCE

Given the important role of data across many areas, this workshop is expected to attract participation of researchers and practitioners from diverse disciplines. These include participants from relevant computer science areas such as databases, information retrieval, human computation and crowdsourcing, human computer interaction, machine learning, and other fields of AI such as natural language processing and computer vision. Besides, the interdisciplinary discussion on human factors and issues relating to responsible data curation, we hope to potentially attract participants with interest and expertise in related areas such as social science, psychology, ethics and philosophy.

3 WORKSHOP RELEVANCE

The challenge of data curation is relevant across ares in computing including artificial intelligence, search and discovery, data mining, and database systems. The quality of data considered in machine learning pipelines is critical for the quality of the output. A lot of time and effort is currently spent both in industry and academia on making sure the best possible data is used as input to data-driven systems. This workshop aims at providing a forum to share experiences and methods that help addressing this significant cost of information and knowledge management systems. Several topics in this year CIKM call for papers relate to this workshop focus, including, for example, 'Data and information acquisition and preprocessing', 'Integration and aggregation', 'Special data processing', 'Users and interfaces for information and data systems', and 'Crowdsourcing'.

4 RELATED WORKSHOPS

Our workshop is related to many other workshops but fills a significant void. In the following, we describe some recent, wide-reaching workshops in the machine learning, human computation, database, and human-computer interaction communities.

The *Data-Centric AI* at Neurips 2021 is a recent effort that brings to the fore the concept of Data-Centeredness to the ML and AI communities. It notably accepted 80 papers and had hundreds of participants, signifying the wide acceptance of the data-centric AI concept. The workshop emphasizes the importance of data quality (especially under a limited amount of labeled data), sharing a

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CIKM '22, October 17–21, 2022, Atlanta, GA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557498

similar scientific problem as in our workshop. The Data-Centric AI workshop, however, has a strong focus on algorithmic methods for working with limited labeled data and improving label efficiency. Our workshop differs significantly in the approach we take, focusing on humans in the loop. Note that the Data-Centric AI workshop was built on a series of workshops focusing on the role of data in AI including such as *Data Excellence* at HCOMP, *Meta-Eval 2020* at AAAI that all share similar goals.

The *Human-Centered Explainable AI* workshop at CHI 2022 advocates the central role of humans in the formulation of technological goals and evaluation of the methods or tools developed. In the HCI community, human- or user-centeredness is a well-accepted concept. Yet the discussions often consider the role of humans as stakeholders in the usage or application of computational systems. In our workshop, we instead focus on the roles of humans being the computational agent in the data curation process (not necessarily limited to explainable AI).

The *Databases and Machine Learning (DBML)* workshop at ICDE 2022 concentrates the discussion on the synergy between the database and ML: how the ML pipeline especially the data preparation pipeline can benefit from data management techniques, and how data-driven approaches such as machine learning can enhance database systems. The workshop is related to ours in the sense that data quality is a long-standing problem in database research and this workshop further connects to machine learning that we also consider in our workshop as an application but also as a method to support human-in-the-loop data curation. The focus of our workshop on humans in the loop is however not discussed in DBML.

The *Subjectivity, Ambiguity and Disagreement in Crowdsourcing (SAD)* workshop at TheWebConf (WWW) 2019 focuses its theme on capitalizing on the uncertainty in human work coming from the subjectivity, ambiguity, and disagreement. As such, the workshop focuses on the human factors in computation that applies to broader research areas including social science such as "communication science, law, and political science". Data problem for AI and advanced analytics was not explicitly addressed as our workshop does.

5 PROGRAM FORMAT

Our workshop will be held mainly in-person. All the three organizers will be able to attend CIKM and organize the workshop in-person. To drive fruitful discussions around the exciting topic of human-in-the-loop data curation, the workshop will feature a diverse set of joint activities by participants. The planed activities are categorized into the following three parts:

- **Part 1** features plenary sessions, including the keynotes, invited talks, and panel.
- **Part 2** features selected presentations from speakers whose papers are peer-reviewed and who attend in person.
- **Part 3** features lightning talks for extended abstracts that are not formally peer-reviewed.

Part 1 and 2 will be held in-person, whereas part 3 allows remote talks to accommodate contributors who cannot attend in-person.

6 PROGRAM COMMITTEE

The program committee is composed of a set of scholars and scientists from diverse backgrounds relevant to our topic.

Name	Affiliation (Country)
Fabio Casati	Servicenow Inc. (USA)
Matt Lease	UT Austin, Amazon (USA)
Marco Brambilla	Politecnico di Milano (Italy)
Jahna Otterbacher	Open University of Cyprus (Cyprus)
Hailong Sun	Beihang University (China)
Jie Zhang	Nanyang Technological University (Singapore)

7 PARTICIPATION AND SELECTION PROCESS

Paper submissions that will be invited for in-person presentations will be reviewed in a peer-review process by at least two members of the program committee. At least one author of each paper will need to register for the conference and attend the workshop to present the paper. Submissions for the lightning talks will be checked by the organisers and up to 12 will be accepted.

8 ORGANIZERS

Gianluca Demartini is an Associate Professor in Data Science at the University of Queensland, Australia. His main research interests include Information Retrieval, Semantic Web, and Human Computation. He received Best Paper awards at AAAI HCOMP in 2018, at ECIR in 2016 and 2020. He has published more than 150 peer-reviewed scientific publications including papers at major venues such as TheWebConf, ACM SIGIR, VLDBJ, ISWC, and ACM CHI. He is an ACM Senior Member.

Shazia Sadiq is a Professor of Data Science, Director of the ARC Training Centre for Information Resilience at The University of Queensland. Her research aims at developing innovative solutions for data management including data quality, data governance, risk and compliance, efficient workflow systems, and advanced analytical solutions to complex business and social problems. She has published 200 peer-reviewed publications in high ranking venues within information systems and computer science such as VLDB Journal, TKDE, Information Systems Journal, World Wide Web Journal, International Conference on Business Process Management (BPM), CAiSE, ACM SIGMOD, International Conference on Information Systems (ICIS), ER and IEEE ICDE.

Jie Yang is Assistant Professor in the Web Information Systems group at TU Delft. He co-leads the *Kappa* research line on Crowd Computing & Human-Centered AI[1] at the WIS group and the Delft AI Lab Design@Scale[2]. His current research focuses on human-in-the-loop approaches for reliable and trustworthy machine learning (ML). He has published more than 60 papers in leading venues of information systems and AI such as The Web Conference/WWW, CHI, SIGIR, CHIIR, HCOMP, AAAI, IJCAI, CIKM, RecSys, TKDE, etc. His work has received the Best Paper Award nomination at the Web Conference (2022), the Best Paper Awards at the 28th ACM HT Conference (2017) and at the International Workshop on Crowd Work (2015), and the 1st Prize Blue Sky Idea and the Best Demo Award, both at AAAI HCOMP (2021).

Acknowledgments. This work is partially supported by an ARC Discovery Project (Grant No. DP190102141), by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

[1] https://www.wis.ewi.tudelft.nl/crowd-computing
[2] https://www.tudelft.nl/en/ai/design-at-scale-lab