Report on the 1st Workshop on Human-in-the-loop Data Curation (HIL-DC 2022) at CIKM 2022

Gianluca Demartini
The University of Queensland
Australia
demartini@acm.org

Jie Yang
TU Delft
Nederlands
J.Yang-3@tudelft.nl

Shazia Sadiq
The University of Queensland
Australia
shazia@itee.uq.edu.au

Abstract

We report on the First Workshop on Human-in-the-loop Data Curation (HIL-DC), which was co-located with the ACM International Conference on Information and Knowledge Management (CIKM) 2022. 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. Current approaches include manual, automated, and hybrid human-machine methods to data curation. However, this topic remains relatively unstudied, so our main aim for organizing this workshop was to bring together a group of people from both industry and academia with an interest in the topic, in order to arrive at a shared roadmap for the future. Through a program that included two keynotes, seven peer-reviewed papers, and six lightening talks, we have made initial steps towards a common understanding and shared research agenda for this timely and important topic.

Date: 21 October, 2022.

Website: https://hilworkshops.github.io/hil-dc2022/.

1 Introduction

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.

In the first workshop on Human-in-the-loop Data Curation (**HIL-DC**), which was held as an hybrid event on October 21, 2022 as part of the 31st ACM International Conference on Information and Knowledge Management (**CIKM 2022**) we had the opportunity to explore the interdisciplinary overlap between manual, automated, and hybrid human-machine methods of data curation. The goal of our workshop was to start a meaningful, long-lasting dialogue spanning researchers across a wide variety of backgrounds and communities on a timely and important but under-explored topic.

2 Workshop Aims and Scope

The scope of the workshop included the following themes:

- Quality control for crowdsourced data curation.
- Data worker incentivization and engagement, including techniques from citizen science and collective intelligence.
- Expertise finding and engagement for data curation.
- Supporting crowd workers and experts in data task completion.
- Supporting data curation task design for data requesters.
- Collaborative data work among humans and between humans and AI.
- Human studies into the transparency, reliability, and biases in manual and hybrid data curation.
- Interaction techniques for manual, collaborative, and hybrid human-machine data curation, e.g., conversational interfaces.
- Database and machine learning techniques for supporting large-scale and hybrid data curation.
- Human intervention in data cascades and machine learning lifecycle management.
- Benchmarks in machine learning, AI, and related areas.
- Privacy and security issues of data quality, e.g., data poisoning attacks.

3 Workshop Contributions

The HIL-DC workshop aimed to start a conversation around a research topic where a lot of practice exists but where there are also limited shared principles and standards. To this end, we assembled a diverse group of around 30 participants across industry and academia. We discussed data curation strategies for diverse data types, including text, multimedia, and structured data. In total, the full-day workshop comprised two keynote speakers, seven accepted papers, as well as short lightening talks with which participants were invited to contribute.

3.1 Keynotes

Our keynotes were delivered by Prof. Abraham Bernstein from the University of Zurich and Dr. Ujwal Gadiraju from the Delft University of Technology:

- **Towards a Collaboration between Humans and Machines for Data Curation and Analysis, by *Abraham Bernstein*.** Our first keynote looked at several examples of data curation performed by hybrid human-machine workflows. Bernstein first illustrated that the human-in-the-loop crowdsourcing paradigm has led to surprisingly good results for tasks that have been assumed to be in the realm of highly specialized professionals. He then examined where the human-in-the-loop paradigm might fail both due to the structure and inherent properties of data curation and analysis tasks. Hence, he argued that we need to explore various approaches to combining the complementing abilities of humans and machines to capture the rationale and improve the performance of data curation and analysis tasks.
- **Human-Centered AI: A Crowd Computing Perspective, by *Ujwal Gadiraju*.** The unprecedented rise in the adoption of machine learning models and artificial intelligence techniques – alongside automation in many contexts – is concomitant with the shortcomings of such technology. There is an increasing body of work that has unearthed concerns regarding the robustness, interpretability, usability, trustworthiness, and explainability of complex machine learning models being used in the real world. Crowd computing offers a viable means to leverage human intelligence at scale for data creation, enrichment, and interpretation, demonstrating a great potential to improve the performance of AI systems and increase the adoption of AI in general. Our second keynote discussed opportunities in crowd computing to propel better AI technology, and argued that to make such progress, fundamental problems need to be tackled from both the computational and interactional standpoints. Understanding the role of trust is pivotal in shaping better human-AI interactions and ensuring truly beneficial AI-assisted decision-making. Gadiraju also shed light on the research needed to help pave a future where humans can benefit by working seamlessly with AI systems and relying on them appropriately.

3.2 Accepted papers

We solicited regular paper submissions presenting existing data curation efforts or frameworks. Submissions were peer-reviewed by members of the workshop program committee. The workshop proceedings including accepted manuscripts are published by CEUR-WS jointly with the others workshops at CIKM 2022. After the review process, we accepted seven submissions that we grouped in three sessions: (1) human-in-the-loop data curation methods, (2) natural language processing, and (3) multimedia.

3.2.1 Methods

- **HITL IRL: 12 Reflections on Expertise Finding and Engagement for a Large Data Curation Team, by *Brendan Coon*.** Abstract: As ML and AI increasingly shape product development, the need for a rigorous humans-in-the-loop approach for quality control increases in importance. Impactful Data Curation teams are responsible for understanding

and assessing the quality of the training data feeding into models and algorithms, and are able to package their evaluations in a consumable and actionable format. This paper covers some of the necessary steps to build a successful Data Curation team that can continuously deliver value, even as your core business or academic use case evolves. By providing an overview of what has worked during my 9 years on the team, I aim to provide an essential guide to building a new team or improve an existing one. My contention is that the unique perspective contained in this paper is advice that can help several disciplines that might be looking after a Data Curation team as part of their remit—researchers, ML engineers, product managers—get high-integrity data and algorithm evaluations from the experts they engage. Building and maintaining a Data Curation team will directly impact any product team's ability to "identify issues with usability and comprehensibility associated most closely with content quality and with the user experience." It is important that you find the right people and retain them—this paper lays out how to do both. Some key takeaways the reader might acquire from this paper are how to find and identify the right experts, how to support and work with those experts, and how to retain and engage those experts. They are mostly pulled from my experience in a business environment, but can also apply to an academic one.

- **Developing a Noise-Aware AI System for Change Risk Assessment with Minimal Human Intervention, by *Subhadip Paul, Anirban Chatterjee, Binay Gupta and Kunal Banerjee.*** Abstract: Introducing changes to a system in production may sometimes result in failures, and eventual revenue loss, for any industry. Therefore, it is important to monitor the "risk" that each such change request may present. Change risk assessment is a sub-field in operations management that deals with this problem in a systematic manner. However, a manual or even a human-centered AI system may find it challenging to meet the scaling demands for a big industry. Accordingly, an automated system for change risk assessment is highly desired. There are a few commercial solutions available to address this problem but those solutions lack the ability to deal with highly noisy data, which is quite a possibility for such systems. There are literature which proposed methods to integrate the feedback of domain experts into the training process of a machine learning model to deal with noisy data. Even though some of these methods produced decent risk prediction accuracy of the model but such an arrangement to collect feedback from the domain experts continuously has practical challenges due to the limitation in bandwidth and availability of the domain experts at times. Therefore, as part of this work, we explore a way to take the transition from a human-centered AI system to a near-autonomous AI system, which minimizes the need of intervention of domain experts without compromising with the prediction accuracy of the model. Initial experiments with the proposed AI system exhibit 10% improvement in risk prediction accuracy in comparison with the baseline which was trained by integrating the feedback of domain experts in the training process.

3.2.2 Natural Language Processin.

- **Knowledge Management System with NLP-Assisted Annotations: A Brief Survey and Outlook, by *Baihan Lin.*** Abstract: Knowledge management systems are in high demand for industrial researchers, chemical or research enterprises, or evidence-based

decision making. However, existing systems have limitations in categorizing and organizing paper insights or relationships. Traditional databases are usually disjoint with logging systems, which limit its utility in generating concise, collated overviews. In this work, we briefly survey existing approaches of this problem space and propose a unified framework that utilizes relational databases to log hierarchical information to facilitate the research and writing process, or generate useful knowledge from references or insights from connected concepts. This framework of knowledge management system enables novel functionalities encompassing improved hierarchical notetaking, AI-assisted brainstorming, and multi-directional relationships. Potential applications include managing inventories and changes for manufacture or research enterprises, or generating analytic reports with evidence-based decision making.

- **Creating a framework for a Benchmark Religion Dataset, by *Deepa Muralidhar and Ashwin Ashok.*** Abstract: Language Models (LM) such as OpenAI's GPT series have made significant progress in generating natural language text in the last few years. The model takes a prompt (textual data) as input and generates text as output that represent the most probable sequence of words matching the prompt's context and pattern. Our preliminary investigations revealed bias but did not give us enough evidence for the cause of the bias. Our first goal is, therefore, to build a benchmark dataset on various religions for evaluating the bias exhibited in the LM's towards religion. We envision that our conceptual method of creating a dataset and developing a bias rating mechanism can serve as a fundamental tool to establish a process to find explanations for the bias. We hypothesize that comparing the bias indicator value (BIV) generated for one religion against another can give us enough information to provide a holistic bias rating for the text generated.

3.2.3 Multimedia

- **A Human-ML Collaboration Framework for Improving Video Content Reviews, by *Meghana Deodhar, Xiao Ma, Yixin Cai, Alex Koes, Jilin Chen and Alex Beutel.* Best paper award** Abstract: We deal with the problem of localized in-video taxonomic human annotation in the video content moderation domain, where the goal is to identify video segments that violate granular policies, e.g., community guidelines in an online video platform. High quality human labeling is critical for enforcement in content moderation. This is challenging due to the problem of information overload - raters need to apply a large taxonomy of granular policy violations with subjective definitions within a limited review duration to relatively long videos. Our key contribution is a novel human-machine learning (ML) collaboration framework aimed at maximizing the quality and efficiency of human decisions in this setting - human labels are used to train segment-level models, the predictions of which are displayed as "hints" to human raters, indicating probable regions of the video with specific policy violations. The human verified/corrected segment labels can help refine the model further, hence creating a human-ML positive feedback loop. Experiments show both significantly improved human video moderation decision quality, and efficiency through more granular annotations submitted within the same review duration, which enable a 5-8% AUC improvement in the policy violation models.
- **From Fat Deposits to Floating Forests: Cross-Domain Transfer Learning using PatchGAN-based Segmentation Model, by *Kameswara Mantha, Ramanaku-***

mar Sankar, Yuping Zheng, Lucy Fortson, Thomas Pengo, Douglas Mashek, Mark Sanders, Trace Christensen, Jeffrey Salisbury, Laura Trouille, Jarrett Byrnes, Isaac Rosenthal, Henry Housekeeper and Kyle Cavanaugh. Abstract: Many scientific domains gather sufficient labels to train machine algorithms through human-in-the-loop techniques provided by the Zooniverse.org citizen science platform. As the range of projects, task types and data rates increase, acceleration of model training is of paramount concern to focus volunteer effort where most needed. The application of Transfer Learning (TL) between Zooniverse projects holds promise as a solution. However, understanding the effectiveness of TL approaches that pretrain on large-scale generic image sets vs. images with similar characteristics possibly from similar tasks is an open challenge. We apply a generative segmentation model on two Zooniverse project-based data sets: (1) to identify fat deposits in breast cancer images (FatChecker; FC) and (2) the identification of kelp beds in satellite images (Floating Forests; FF) through transfer learning from the first project. We compare and contrast its performance with a TL model based on the COCO image set, and subsequently with baseline counterparts. We find that both the FC and COCO TL models perform better than the baseline cases when using ¿75% of the original training sample size. The COCO-based TL model generally performs better than the FC-based one, likely due to its generalized features. Our investigations provide important insights into usage of TL approaches on multi-domain data hosted across different Zooniverse projects, enabling future projects to accelerate task completion.

- **Human-AI Collaboration for Improving the Identification of Cars for Autonomous Driving, by *Edwin Gamboa, Jose Alejandro Libreros Montaño, Dan Dubiner and Matthias Hirth.*** Abstract: Large and high-curated training data is required for Artificial Intelligence (AI) models to perform robustly and reliably. However, training data is scarce since its production normally requires manual expert annotation, which limits scalability. Crowdsourced micro-tasking can help to overcome this challenge, as it offers access to a global workforce that might enable high-scalable annotation of visual data in a cost-time effective way. Therefore, we aim to develop a workflow based on Human-AI collaboration that shall enable large-scale annotations of image data for autonomous driving systems. In this paper, we present the first steps towards this goal, in particular, a Human-AI approach for identifying cars. We assess the feasibility of this collaboration via three scenarios, each one representing different traffic and weather conditions. We found that crowdworkers improved the AI's work by identifying more than 62% of the missing cars. Crowdworkers' contribution was key in challenging situations in which identifying a car depended on context.

In addition to fully peer-reviewed papers, we had a call for lightening talk submissions and ended up also including the following talks into the workshop program interleaved with accepted paper presentations:

- Improving Labeling Through Social Science Insights, by Stephanie Eckman, Jacob Beck, Rob Chew and Frauke Kreuter.
- Human-AI Complex Task Planning, by Sepideh Nikookar.
- PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data, by Shubhanshu Mishra and Jana Diesner.

- A Paradigm to Put Back the User into the AutoML Loop through Natural Language, by Sara Pidò and Pietro Pinoli.
- Help Me Help You - A Mixed-Initiative Approach To Explore Book-length Documents, by Bipasha Banerjee, Palakh Mignonne Jude, William A. Ingram, Kurt Luther and Edward A. Fox.

4 Roadmap and Shared Research Agenda

Based on the submissions, presentations, and discussions during the workshop, we summarize next the main topics of discussion and the key open research challenges in this area. We follow the Information Resilience framework [Sadiq et al., 2022] to organize the main topics discussed during the workshop, aiming at summarising the main research directions that have been presented and discussed by the workshop participants. The key functions of robust data pipelines that have been identified by Sadiq et al. [2022] include i) responsible use of data assets, ii) data curation at scale, iii) algorithmic transparency, iv) trusted data partnerships, v) agility in value creation from data.

During the workshop several authors discussed open research topics that need to be addressed by the community. For example, the importance and the processes required to collect robust data annotations (e.g., Coon from Spotify). Another recurrent topic was the impact of noisy labels on machine learning models related to the need for data curation and transparent algorithms. The support that algorithms can provide to human decision making is described by the paper by Lin. The presence and propagation of bias was a common issue in data pipelines mentioned by participants. A good example of this is the work by Muralidhar and Ashok. The process of human annotation was also discussed by Xiao Ma from Google in the context of YouTube content moderation. They showcased how an hybrid intelligence pipeline can achieve effective and scalable content moderation. Similarly, Gamboa discussed the scenario of autonomous cars and how hybrid intelligence systems can provide annotations that can help algorithms learning from otherwise scarce training data.

In summary, during the workshop the key emerging research topics related to the data annotation process, the role of human intelligence in otherwise fully automated data pipelines, and the presence of bias in human annotations and machine learning systems. An important question in such context is who should do what. This also relates to the work by Nikookar on task planning and on deciding whether a specific data point should be given to or a decision should be taken by a human or by an algorithm. The breadth of the workshop has also confirmed how these topics relate to any type of digital content, being it textual, visual, or numerical.

5 Conclusion

The HIL-DC 2022 workshop was a successful event that kick-started an important cross-community conversation on data curation approaches and standards and the open research questions that relate to making data better by leveraging a combination of human and machine intelligence. We received input from researchers and practitioners focused on various aspects of data curation for different data formats and data quality issues. Data curation and content moderation has

been prevalently been discussed by industry speakers. Workshop participants were well balanced between in-person and remote attendance with a good representation from industry (including participants from Google, Spotify, Walmart, Twitter, and Adobe) and academia (including participants from MIT, Virginia Tech, and Georgia State). We thank CIKM 2022 for their support in organizing the event, especially given the complex hybrid format that required an advanced A/V setup to be put in place. We hope to expand upon the workshop in the following years.

Acknowledgments

The workshop was partially supported by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

References

Shazia Sadiq, Amir Aryani, Gianluca Demartini, Wen Hua, Marta Indulska, Andrew Burton-Jones, Hassan Khosravi, Diana Benavides-Prado, Timos Sellis, Ida Someh, et al. Information resilience: the nexus of responsible and agile approaches to information use. *The VLDB Journal*, pages 1–26, 2022.