HILDA 2023

Workshop on Human-In-the-Loop Data Analytics

Co-located with SIGMOD 2023 (18 June 2023, Seattle, WA, USA)

Follow @hildaworkshop Post #hilda2023

Location: Hyatt Regency Bellevue hotel, Room **Evergreen BC**

Past workshops: HILDA 2022 | HILDA 2020 | HILDA 2019 | HILDA 2018 | HILDA 2017 | HILDA 2016

HILDA brings together researchers and practitioners to exchange ideas and results on **human-data interaction**. It explores how data management and analysis can be made more effective when taking into account the people who design and build these processes as well as those who are impacted by their results.

In HILDA 2022, we implemented a **mentoring program** (inspired by workshops such as PLATEAU) and are continuing it this year. Our focus is on promising and early-stage research, with a core component of the program being that each paper is assigned a **mentor**. More details on the process are below.

The theme for this edition of the workshop is **commodifying human-in-the-loop data analytics**, however, the workshop is not limited to this theme and other topics are also of interest. Despite significant progress in human-in-the-loop approaches in recent years, many systems are still not ready for end-user consumption. We aim to provide guidelines and best practices for creating HILDA-as-a-Service systems that can be easily integrated into larger data ecosystems and used by different types of users in a plug-and-play fashion.

PROGRAM

The times in the following schedule are all in Pacific time. The schedule is not finalized and may change. Please check it again closer to the event.

8:45 Opening Remarks

9:00 **Keynote by Bill Howe: Curation as Programming: AI, Data Management, and Mediated Knowledge Interaction**

9:45 **Keynote by Ana Crisan: Scaling data-driven Decision-making through Human-AI Interaction**

10:30 *Break*

Session Chair: Dominik Moritz

11:00 **Facilitating Dependency Exploration in Computational Notebooks**
Colin Brown, David Koop, Hamed Alhoori

11:20 **VALUE: Visual Analytics-driven Linked Data Utility Evaluation**
Kaustav Bhattacharjee, Aritra Dasgupta

11:40 **SliceLens: Guided Exploration of Machine Learning Datasets**
Daniel Kerrigan, Enrico Bertini

12:00 **DIG: The Data Interface Grammar**
Yiru Chen, Jeffrey Tao, Eugene Wu

12:20 **(Short Paper) Visualizing a Tabular Data Repository to Facilitate Descriptive Tag Augmentation for New Tables**
Jianhao Cao, Tamara Munzner, Rachel Pottinger

12:30 *Lunch Break*

Session Chair: Sudeepa Roy

13:30 **Keynote by Yunyao Li: Building, Growing and Serving Large Knowledge Graphs with Human-in-the-Loop**

14:15 **Overlay Spreadsheets**
Oliver A Kennedy, Boris Glavic, Michael Brachmann

14:35 **A Human-in-the-loop Workflow for Multi-Factorial Sensitivity Analysis of Algorithmic Rankers**
Jun Yuan, Aritra Dasgupta

15:00 *Break*

Session Chair: Behrooz Omidvar-Tehrani

15:30 **Camera-First Form Filling: Reducing the Friction in Climate Hazard Reporting**
Kristina Wolf, Dominik K Winecki, Arnab Nandi

15:50 **Raven: Accelerating Execution of Iterative Data Processing Workflows by Reusing Results of Previous Equivalent Versions**
Sadeem Alsudais, Avinash Kumar, Chen Li

16:10 **Aggregation Consistency Errors in Semantic Layers and How to Avoid Them**
Zezhou Huang, Pavan Kalyan Damalapati, Eugene Wu

16:30 **(Short Paper) Interactive Data Cleaning for Real-Time Streaming Applications**
Timo Räth, Kai-Uwe Sattler, Ngozichukwuka Onah

16:40 **(Short Paper) Approximate Query Answering over Open Data**
Mengqi Zhang, Pranay Mundra, Chukwubuikem Chikweze, Fatemeh Nargesian, Gerhard Weikum

16:50 **(Short Paper) Data Makes Better Data Scientists**
Jinjin Zhao, Avigdor Gal, Sanjay Krishnan

17:00 Closing Remarks

HILDA 2023 KEYNOTE TALKS

Our exciting program will feature the following invited keynote speakers to talk about the challenges of human-data interaction.

Bill Howe, Associate Professor, University of Washington

Title: Curation as Programming: AI, Data Management, and Mediated Knowledge Interaction

Abstract: AI advances have been concentrated in curation-on-read settings: LLMs are trained on massive, weakly curated convenience samples of the internet, while the output is assumed to be subject to careful human review and accountability on a per-instance basis. This regime shifts all responsibility to the end user to identify errors, biases, and compliance issues (e.g., intellectual property violations). There do exist successful AI applications in curation-on-write settings, typically in science: models trained on precise, objectively correct input in order to produce precise, objectively correct output. For example, popular deep learning architectures are poised to outperform physics-based models to predict the weather, simultaneously learning the physics and the parameters directly from observations, without requiring fluid dynamics to be explicitly programmed. We are studying enabling technologies to reduce the cost of developing AI systems in these curation-on-write settings, characterized by limited data, complex multi-modal features, and ambiguous or conflicting labels. As methods continue to be commoditized, costs are driven by finding and organizing training and evaluation data. This perspective of “curation as programming” is an opportunity to design new tooling, particularly in HCI, to empower domain experts to build safer AI systems for specialized tasks in high-expertise settings. I'll describe some projects my group is pursuing in this space including synthesizing data in urban settings, identifying speakers and agenda items in local council meetings, and extracting information from legal documents. I'll describe some general technical questions we encounter in these settings, including how best to use expert-provided ontologies and unlearning specific biases during fine-tuning.

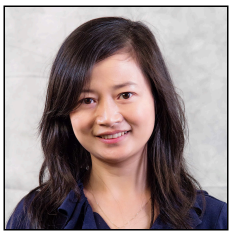
Speaker Bio: Bill Howe is Associate Professor in the Information School and Adjunct Associate Professor in the Allen School of Computer Science & Engineering and the Department of Electrical Engineering. His research interests are in data management, machine learning, and visualization, particularly as applied in the physical and social sciences. As Founding Associate Director of the UW eScience Institute, Dr. Howe played a leadership role in the Moore-Sloan Data Science Environment program through a $32.8 million grant awarded jointly to UW, NYU, and UC Berkeley, and founded UW's Data Science for Social Good Program. With support from the MacArthur Foundation, NSF, and Microsoft, Howe directs UW's participation in the Cascadia Urban Analytics Cooperative. He founded the UW Data Science Masters Degree, serving as its inaugural Program Chair, and created a first MOOC on data science that attracted over 200,000 students. His research has been featured in the Economist and Nature News, and he has authored award-winning papers in conferences across data management, machine learning, and visualization. He has a Ph.D. in Computer Science from Portland State University and a Bachelor's degree in Industrial & Systems Engineering from Georgia Tech.

Anamari (Ana) Crisan, Lead Research Scientist, Tableau

Title: Scaling data-driven Decision-making through Human-AI Interaction

Abstract: Decision-making with data is primarily conducted by professionals lacking formal training in data science, statistics, or machine learning. Aiding these professionals are emerging technologies supported by machine learning (ML) and/or artificial intelligence (AI) techniques that automate many aspects of data work. However, this collaboration between humans and AI/ML technology is far from frictionless and can produce incorrect or misleading results. In this talk I discuss a human-centered approach for ML/AI that aims to address these limitations. I will ground this discussion in human factors studies, techniques, and tools for supporting human-ML/AI interaction (HAI) in the context of data discovery and analysis. I will contrast existing approaches along the dimensions of user, data, and model centric paradigms. Finally, I will propose a forward looking research approach to developing trustworthy and responsible HAI techniques for a broad analyst population.

Speaker Bio: Anamaria (Ana) Crisan is a Lead Research Scientist at Tableau, where she drives a multidisciplinary research agenda for human-ML/AI interaction that occupies the intersection of applied machine learning, human computer interaction, and data visualization. She completed her Ph.D. in 2019 as a Vanier scholar at the University of British Columbia under the joint supervision of Drs. Tamara Munzner and Jennifer Gardy. Her doctoral research focused on visualization of genomic epidemiological data. Prior to her doctorate, she led research initiatives for personalized cancer care and public health genomics. Her research has appeared in top-tier biomedical journals (Nature, Bioinformatics) as well as conferences of the ACM (CHI, FAccT) and IEEE, has won paper awards (CHI, VDS), contributed to the open source ecosystem, and resulted in patents for ML backed products in use today.



Yunyao Li, Head of Machine Learning, Apple Knowledge Platform

Title: Building, Growing and Serving Large Knowledge Graphs with Human-in-the-Loop

Abstract: The ability to build large-scale knowledge bases that capture and extend the implicit knowledge of human experts is the foundation for many AI systems. We use an ontology-driven approach for the building, growing and serving of such knowledge bases. This approach relies on several well-known building blocks: document conversion, natural language processing, entity resolution, data transformation and fusion. In this talk, I will discuss wide range of real-world challenges related to the building of these blocks and present our work to address these challenges via better human-machine cooperation.

Speaker Bio: Yunyao Li is the Head of Machine Learning, Apple Knowledge Platform, where her team builds the next-generation machine learning solutions to help power features such as Siri and Spotlight. Previously she was a Distinguished Research Staff Member and Senior Research Manager at IBM Research - Almaden, leading the building and delivery of core language technologies to over 20 IBM products and solutions. She also co-led the IBM-Stanford HAI partnership. Her technical contributions span the areas of natural language processing (NLP) , data management, information retrieval, and human computer interaction. In these areas, she has published over 100 publications, including a

book, and had more than 50 patent granted/filed. She is an ACM Distinguished Member. She is a member of the New Voices program of the US National Academies. She was an IBM Technology Academy Member and a Master Inventor. She has served the NLP and database communities with distinction and regularly serves as organizer and senior committee member for top conferences and editorial board member. She is currently an elected member of NAACL Executive Board. She received her undergraduate degrees from Tsinghua University, and her masters and Ph.D. in Computer Science from the University of Michigan - Ann Arbor.

WHAT TO SUBMIT

We encourage both standard research papers and more unusual works—for instance papers that describe in-progress work, reports on experiences, question accepted wisdom, raise open problems, or propose speculative new approaches. A HILDA submission should describe *work or perspectives* that will lead to interesting discussions at the workshop or that the authors want feedback on.

We welcome work that proposes innovations in design to improve the way people can work with data management systems, as well as work that studies empirically how humans interact with existing systems. We welcome research that comes from the traditions of the database systems community, and also reports on industry activities, and research on data topics from communities that study people and organizations. A sample of topics that are in the spirit of this workshop include, but are not limited to:

- novel query interfaces,
- interactive query refinement,
- data exploration and analysis,
- data visualization,
- human-assisted data integration and cleaning,
- perception-aware data processing,
- database systems designed for highly interactive use cases,
- empirical studies of database use,
- evaluating and ensuring fairness in data-driven decision making processes
- understanding the outcomes of processes through provenance and explanations
- interactive debugging of complex data systems
- crowd-powered data infrastructure, etc.

Submissions can also examine any of the above topics from an application or domain perspective.

HILDA is a forum where people from multiple communities engage with one another's ideas. We are keen to have submissions that present initial ideas and visions, just as much as reports on early results, or reflections on completed projects.

The workshop will focus on discussion and interaction, rather than static presentations of what is in the paper.

REVIEW AND MENTORSHIP PROCESS

HILDA reviews are single blind. All submitted papers will be reviewed by at least three reviewers who will determine the fit of the work for HILDA's unique mentorship process this year, the quality of the work, and its potential for future research.

Every accepted paper will be assigned a mentor who will engage with the authors providing constructive feedback through one-on-one, virtual, discussions. We hope that the authors will work closely with their mentors to improve the substance and direction of their work.

Authors and mentors can withdraw without repercussions due to unforeseen conflicts. In such situations, the program chairs will try to find another suitable mentor.

All accepted papers will have the chance to demonstrate the work during the workshop's demonstration session. For papers selected for presentation at the conference, the mentors will also introduce and contextualize the work and enable feedback from the attendees.

Final camera-ready submissions will be solicited for a selection of the accepted papers after the workshop allowing authors to integrate feedback from discussions with the mentor and workshop participants.

SUBMISSION

Authors are invited to submit papers between four and six pages in length excluding references and using the standard SIGMOD paper formatting template Submissions should reflect the current state of the research work but also include a section on limitations and challenges that they wish to receive feedback from their mentors and the HILDA community on.

All submissions must follow the latest ACM paper format with 10pt font size.

Submission website: https://cmt3.research.microsoft.com/HILDA2023

PROCEEDINGS

We will provide links to accepted papers in the program here as well as publish them for a year through the ACM DL.

IMPORTANT DATES

- Workshop Date: June 18, 2023
- Submissions: ~~March 30~~ April 6, 2023 AOE
- Notification of outcome: April 30, 2023 (Tentative)
- Camera-ready due: July 1, 2023 (after the workshop)

WORKSHOP CHAIRS

- Dominik Moritz (Carnegie Mellon University)
- Behrooz Omidvar-Tehrani (AWS AI Labs)
- Sudeepa Roy (Duke University)

MENTORS

- Amit Somech (Bar-Ilan University)
- Anoop Deoras (AWS AI Labs)
- Brit Youngmann (CSAIL MIT)
- Cagatay Demiralp (Sigma Computing)
- Chengkai Li (The University of Texas at Arlington)
- Dixin Tang (University of California, Berkeley)
- El Kindi Rezig (MIT)
- Eugene Wu (Columbia University)
- Jean-Daniel Fekete (Inria)
- Kanit Wongsuphasawat (Databricks)
- Kurt Stockinger (ZHAW Zurich University of Applied Sciences)
- Oliver Kennedy (University at Buffalo, SUNY)
- Protiva Rahman (Vanderbilt University Medical Center)

- Roee Shraga (Northeastern University)
- Sainyam Galhotra (University of Chicago)
- Senjuti Basu Roy (New Jersey Institute of Technology)
- Sepideh Nikookar (New Jersey Institute of Technology)
- Slava Novgorodov (Tel Aviv University)
- Vidya Setlur (Tableau Research)
- Yifei Ma (AWS AI Labs)
- Zhengjie Miao (Duke University)

STEERING COMMITTEE

- Carsten Binnig (TU Darmstadt)
- Juliana Freire (New York University)
- Joseph M. Hellerstein (University of California, Berkeley)
- Aditya Parameswaran (University of California, Berkeley)

CONTACT

For questions, please email the workshop chairs directly.

FOLLOW US

Join us on Twitter.