Editorial: Artificial Intelligence for Data Discovery and Reuse Demands Healthy Data Ecosystem and Community Efforts

Huajin Wang
huajinw@cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Keith Webster
kwebster@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

ABSTRACT

There is great value embedded in reusing scientific data for secondary discoveries. However, it is challenging to find and reuse the large amount of existing scientific data distributed across the web and data repositories. Some of the challenges reside in the volume and complexity of scientific data, others pertain to the current practices and workflow of research data management. AIDR 2019 (Artificial Intelligence for Data Discovery and Reuse) is a new conference that brings together researchers across a broad range of disciplines, computer scientists, tool developers, data providers, and data curators, to share innovative solutions that apply artificial intelligence to scientific data discovery and reuse, and discuss how various stakeholders work together to create a health data ecosystem. This editorial summarizes the main themes and takeaways from the inaugural AIDR conference.

CCS CONCEPTS

• **Applied computing** → **Digital libraries and archives**; • **General and reference** → *General conference proceedings*;

KEYWORDS

data reuse, data discovery, research data management, artificial intelligence, metadata

1 INTRODUCTION

On May 13-15, 2019, Carnegie Mellon University (CMU) Libraries and Pittsburgh Supercomputing Center (PSC) co-hosted a 2.5-day conference, AIDR 2019: Artificial Intelligence for Data Discovery and Reuse[3]. The theme of the conference is to bring all stakeholders together to share practices and discuss solutions that leverage advances in machine learning (ML) and artificial intelligence (AI) to solve challenging issue in finding and reusing the large amount of existing scientific data and make new discoveries from these valuable data. Approximately 150 participants from 10 countries and 65 institutions and organizations joined the conference. Participants came from academia, industry, non-profit organizations and government, that included researchers across many disciplines, computer scientists, data professionals, and infrastructure providers.

2 BACKGROUND

The volume and complexity of scientific data has increased exponentially in the past decade. Even though these data are expensive to produce and extremely valuable as assets for new discoveries, it is difficult to reuse them and extract information needed. With increasing mandates from funding agencies and publishers, recommendations from organizations and consortia, and rapid emergence of numerous data repositories and tools to facilitate data sharing, there is a positive growth of scientific data being publicly shared more broadly. However, this does not make finding and reusing these shared datasets much easier. It is challenging to find relevant datasets across dispersed institutional, organizational and disciplinary data repositories, each with different database schema and metadata, and even harder to find datasets shard on personal websites and within non-structured web content. Even after a promising dataset is found, it is difficult to understand the data and how the data was generated, evaluate data quality, and integrate with other relevant datasets. Some of these challenges are due to missing documentation and metadata, some lies in the intrinsic heterogeneous nature of the dataset, and others due to lack of consistent data standards and file formats. Facing these challenges, there is an urgent need for ML and AI tools to help researchers find and work with data, especially domain specialists who are not data scientists.

To encourage and provide guidance for data and research output sharing, NSF published a Public Access Plan [2] in 2015, with an ultimate goal to "integrate all publications, data and other products of NSF funding" and broaden access to these findings. In 2018, NSF further announced Dear Colleague Letter (DCL): Advancing Long-term Reuse of Scientific Data, that calls out social and technical infrastructure solutions that support its commitment to public access.

Under this background, CMU Libraries and PSC convened the community from across a broad range of disciplines to discuss solutions that ML and AI may provide to enhance the discoverability and reusability of scientific data.

3 SESSION SUMMARIES

3.1 Opening remarks

The conference opened with Dr. Beth Plale, NSF's program director for Office of Advanced Cyberinfrastructure (OAC) and Public Access. Dr. Plale put the conference into context of NSF's Harnessing the Data Revolution Big Idea, and pointed out that the data revolution requires engaging research community in the development of three intersecting areas: research, educational pathways, and advanced cyberinfrastructure.

Michael McQuade, Vice President of Research at CMU, Keith Webster, Dean of CMU Libraries, and Nick Nystrom, Interim Director of PSC, welcomed participants, and reiterated the importance of the topic from their different perspectives.

3.2 Keynotes

Keynote speaker Tom Mitchell, Interim Dean of CMU's School of Computer Science, pioneer of ML and AI, renowned computer

science professor, talked about his research using ML models to interpret word meanings from dynamic fMRI images of the brain. He pointed out that, there is a big opportunity to jointly analyze multiple types of experimental data from many different experiments, however, hurdles exist due to lack of culture in data sharing, and lack of formal language to specify experiments. He emphasized that the greatest difficulty in cognitive neuroscience is to document every detail of the experiment, and to document in a way that computers can understand, highlighting the value and challenges of scientific data reuse.

Similarly, the second day keynote speaker, Glen de Vries, President and Co-founder of Medidata Solutions, voiced challenges of creating value from clinical trials data due to data standardization issues, independent from data aggregation and AI analysis. As potential solutions, he shared the company's platform that creates synthetic control arms for interpreting uncontrolled trials, and integrating multiple clinical trials data to solve data scarcity issues in precision medicine.

3.3 Session 1: Automation in data curation and metadata generation

Talks in this session addressed one of the most outstanding issues for data discovery and reuse: data curation and creating metadata that is machine readable. One of the shared problems is the large size and heterogeneity of the datasets, including publications data, images, and mixed data. Especially for large-scale astronomy and archaeology images where data is difficult to obtain, reusability of data is especially crucials. Various ML algorithms have been used to automate metadata extraction from these datasets, making downstream data query and analysis possible. However, human and expert knowledge is still an essential piece to ensure the quality of the models.

Cornelia Caragea shared a novel neural learning models to extract keyphrases from large scholarly publications data that integrates citation context. Cheryl Telmer reported Dynamic System Explanation (DySE), a framework that is used to extract information from biomedical publications and to present the extracted knowledge into tabular schema for both humans and machines to understand. Matias Carrasco Kind shared a visualization and classification tool for galaxy images obtained from large astronomical surveys, which enables compression of high dimensional image data into 1-d latent space, to be used for finding images by similarity ranking. Rema Padman demonstrated a patient educational video retrieval system that extracts features from metadata, video, audio and text of YouTube data, and at the same time takes doctor input into consideration to achieve high quality information using a co-training model. Claudia Engel showed a comprehensive project that applies image recognition to large image dataset with incomplete metadata, to capture 25 years of project output from large archaeological sites.

3.4 Session2: Automation in data discovery

The most applauded talk of the session was delivered by Natasha Noy, staff scientist from Google AI, who introduced the Google Dataset Search, a search engine that find datasets across data repositories using metadata embedded in structured web data. Despite representing the giant in technology and AI, Natasha calls out for a healthy data ecosystem, where culture and incentive for data stewardship, and open, non-proprietary data standards are the key ingredients, before AI comes in.

Other talks in this session also shared search engines, platforms and methods, with applications primarily in data discovery from scholarly publications, web archives, and large public databases. These research projects provided some AI-driven solutions for common problems such as finding datasets for data augmentation, reproducibility of data analytics pipelines, and finding documents that are useful and reusable.

Fernando Chirigati described Auctus, a dataset search engine that targets the problem of incomplete or insufficient data, finds datasets that can be joined or unioned from the web, and uses these datasets for data augmentation. Alexander New described a semantalytics framework, that employs application-specific cartridges (subgraphs or modules) to allow reusability and reproducibility when the same heavily used dataset (eg. national public health databases) is used to answer different questions of interest. Shenghui Wang shared a scalable semantic word embedding method that is lightweight and suitable for applications in libraries, for example, efficiently predicting MeSH subject headings. Cornelia Caragea proposed a method that dynamically fuses multiple classification models for PDF document classification, which helps institutions hosting large web archiving collections to automatically determine if a given PDF document is within the scope of a given collection policy. Jian Wu described CiteSeerX, a digital library search engine that maintains an openly accessible big scholarly dataset, and employs a series of ML pipelines (metadata extraction, name disambiguation, subject category classification, etc) to discover reusable datasets from big scholarly data.

3.5 Session 3: Integrating datasets and enabling interoperability

A challenging problem researchers face is that existing datasets often provide only part of the information needed to make conclusions. It is beneficial to integrate and make use of multiple datasets at hand, but at the same time, challenging and time-consuming to do. Speakers from this session discussed the application of AI tools and collaboration platforms in various research areas; topics touched on real-life traffic data, data-intensive animal behavior studies, data augmentation algorithms, and techniques to generate synthetic data for training.

Evgeny Toropov shared their research using domain adaptation algorithms to reuse a small set of well-annotated road images datasets to train AI models for road segmentation in real-life, more heterogeneous road and traffic scenarios. Jiacheng Zhu presented a traffic primitive-based data unification framework that automatically unifies and labels heterogeneous traffic data to allow all data to be labeled and indexed according to traffic scenarios. Daniel Clothiaux demonstrated ML algorithms to integrate large scale font datasets with handwriting datasets into large scale synthetic data, which provides better training results for optical character recognition of typed forms. Xu Fei, a representative from Code Ocean, demonstrated a computational reproducibility cloud platform that

allows users to deposit code, data, and the computational environment as containers and enables reuse of published data and re-run analysis in a browser. Rémi Mégret shared a large scale project to learn honey bee behavior by integrating multiple data types (video recordings, honeybee positions, decoded tags, individual trajectories and behavior events), a project that requires close collaboration between Biologists and Computer Scientists for semi-automated data analysis and pattern extraction.

3.6 Session 4: Biomedical applications

Considering the scale and complexity of data produced in biomedical research, the conference dedicated an extended session on the applications of AI in biomedical data reuse, with topics spanning data integration, data discovery, access and retrieval, and tools and platforms for data analysis, annotation and modeling.

Casey Greene talked about a powerful computational pipeline, multiPLIER, that can reuse large, public genomics data to train ML models that are transferable to much smaller datasets of rare disease samples, and enable cross-dataset and cross-tissue comparisons. Ben Busby showcased a series of projects developed during NCBI hackathons, where prototype software were developed for ML analysis of human genomes, variants, and expression, using a combination of curated NCBI data and user provided data. Sean Davis pointed out that despite the FIAR policy for biomedical research data, FAIR data is not always usable data, and highlighted some scalable data engineering tools in Bioconductor (a collection of R packages), that help to easily and reproducibly collect, transform, explore, and label gene expression data at scale. Irene Kaplow shared her research using Convolutional Neural Network (CNN) to predict tissue specific enhancer sequences of new species from patterns learned from known enhancer sequences, obtained from publicly available genomics data of several known species. Fiona Nielsen presented Repositive, an AI-driven search engine and marketplace that help researchers and contract research organizations to share and find the right genomics data and cancer models, and once again, calls out the community to make easy-to-use tools and fix the incentives for sharing. Nick Nystrom talked about the Human BioMolecular Atlas Program (HuBMAP) that integrates tools and techniques to generate 3D tissue maps by integrating and modeling high-content imaging and omics data to allow FAIR sharing of the HuBMAP data through collaborations among funders, HPC, and research community. Bob Murphy gave a visionary talk on automated sciences, where ML models are continuously improved by small set of experiments, tested retrospectively with small amount of existing data (eg. in publicly available databases), and the model output will in turn guide the "self-driving instruments" to select the next set of experiments to do.

Ethics issues of AI in biomedical research were also addressed. Lisa Parker shared thought-provoking views on privacy and control of private information, arguing that instead of absolute privacy control and informed consent, the right of privacy relies on building a healthy data ecosystem, and on the proper governance of privacy during data collection, use and sharing. Alex London emphasized the ethics of using medical AI systems for diagnosis or intervention, stating that to ensure accountability in decision making, verification rather than explainability is needed to offset the intrinsic uncertainty for medical AI, where outputs should be treated as hypotheses to be verified by randomized controlled trials.

3.7 Session 5: Data security, privacy and algorithmic bias

This session extended the privacy and security conversation around biomedical data from the last session to broader applications, and offered some technical solutions to this very important topic.

Matt Fredrikson described techniques that identify "proxy use" (explicit or inferred use) of personal information in ML models, and provided a "white-box analysis" approach to identify and remove components that lead to biased predictions, a problem that is especially significant for neural networks, algorithms that are known for bias amplification. Andrew Yale presented a synthetic data generator, HealthGAN, that takes a real world medical dataset containing mixed data types to generate synthetic versions of the dataset that can be used in a public environment such as education and research. Michael Ellis talked about using compressed sensing and compressed learning techniques with two-state Markov chain to build a privacy-preserving predictive model for functional magnetic resonance imaging (fMRI) data. Lena Pons discussed the great value in sharing cyber threat information to detect adversarial events, as well as the risks and challenges of sharing cyber threat intelligence data, and further proposed a method using hashing to reduce the risk in such processes.

3.8 Panel Discussions

Two panel discussions took place during the conference. The first day of the conference featured a panel discussion on challenges and opportunities in data reuse using the power of AI. The panel was moderated by Keith Webster, CMU's Dean of the Libraries, and consisted of experts with a wide range of expertise: computational biologist, computer scientist, HPC provider, science policy expert, and AI privacy expert. Discussions touched on broad topics in AI and data reuse, with a few prominent themes emerged.

One of the broadly discussed topics is around the greatest challenges and opportunities for the use of AI in data reuse. Natasha Noy, staff scientist of Google AI and team lead for Google Dataset search, identified that the real challenge in data reuse is not AI, but to foster a community with all stakeholders on board, to provide sufficient incentives for data stewardship, and to build a healthy data ecosystem. This sentiment was shared by all panelists and participants.

When asked about roles that different stakeholders play, Clifford Lynch, Executive Director of the Coalition for Networked Information, stated that funders and universities should play major roles in setting policy and provide incentives for data sharing and to make research reproducible and reusable. Natasha pointed out that for-profit organizations like Google should also play a role in not only advancing technology, but more importantly, advocating for building a healthy data ecosystem. Nick Nystrom, Interim Director of PSC, also highlighted the essential role that high performance computing centers play in providing AI and HPC capacity and hosting datasets and data repositories, and posted a challenging yet prevalent question of where data live over time.

When asked about incentives for data reuse, Casey Greene, Associate Professor at Perelman School of Medicine, University of Pennsylvania, pointed out the fact that not all data will be reused; while we want to set metrics and incentives for data reuse, we should also ensure that people who spend their time and effort sharing data are not punished for the lack of reuse of their data.

Another topic being discussed extensively is around privacy and consent of reusing biomedical data. Alex London, Professor of Ethics and Philosophy at Carnegie Mellon University, pointed out that when reusing data, it is important to evaluate how good the data is, what bias is contained in the data, and posted deeper questions about what aspects of the data are captured, and what information is missing from the data. Others also added that, more data is not necessarily better; when we design experiments and collect patient data, it is important to carefully consider how much data needs to be collected.

At the end of the discussions, the audience and panelists all agreed that there is a lot more to think about on the topics of AI for data discovery and reuse. Data and AI are powerful, but they are not wizards. We need to find ways to establish metrics to measure how much reuse there is, how practical it is to reuse the data, and to set practical goals on what AI can do to facilitate data reuse.

The second panel featured projects from Metro21, an initiative from CMU's Smart Cities Institute led by Executive Director Karen Lightman. Panelists Robert Tamburo, Senior Project Scientist at CMU's Robotics Institute, Laura Meixell Cunniff of City of Pittsburgh, and Bob Gradeck of Western Pennsylvania Regional Data Center, depicted the strong partnerships among the university, government, private and non-profit sectors, who work together to address real-world issues, from transportation to public safety to environmental monitoring. The Metro21 initiative set an example of fostering collaborations among community, research, and data providers to accelerate data reuse and translate data into value.

4 KEY TAKEAWAYS

Data discovery and reuse is a problem shared among different research fields. AIDR convened an audience across a broad range of disciplines, filled a major gap at the intersection of AI and research data management, and stimulated much needed conversations on how to use AI to solve research data issues and collaborate across disciplines and domains. Participants reached the consensus that community building is the key ingredient that comes before any AI can be done. ML and AI is only the end piece, but before we get there, we need to build a healthy data ecosystem, build an interdisciplinary community, build incentives to encourage data sharing and good data practices. A strong sense of an emerging community developed at the meeting.

On the technical end, major types of research data that were discussed include scholarly publications data, image data from various disciplines, health data, and traffic data. Big data appeared to be a growing trend, however, even small data can be challenging to reuse and algorithms using small and noisy data are especially useful for certain applications. Several dataset search engines and platforms were presented, and some shared challenges were discussed, eg., dataset ranking algorithms that take reputation into consideration, linking data with publications, correcting missing or messy metadata. Despite increased complexity of algorithms, human in the loop approaches are still indispensable to define ground truth, provide expert opinion, and provide guidance on more complicated tasks. Incomplete and/or noisy data and the subjectivity of the annotation process make automation tasks technically challenging. There were also intensive conversations around the data privacy and security, especially related to biomedical and health data, and technical, practical, and philosophical solutions were proposed to preserve data privacy and limit algorithmic bias.

5 OUTCOME AND FUTURE PLANNING

Based on the well attended town hall meeting and survey results, participants demand to make AIDR an annual event and to create a mailing list to stay connected as a community (to sign up: aidr-all@lists.andrew.cmu.edu). There was a strong interest to have more presence from industry, especially, a dedicated session for implementing applications from research to industry. There were demands to incorporate broader disciplines and topics into the program, eg., agriculture, health, and psychology, and to extend privacy and ethics discussions to a whole panel. Audience were enthusiastic about adding a workshop day that discuss technical details of, eg., imputation, missing data, and unintended biases. The suggestion to encourage speakers make training data available and reveal datasets used at the end of each talk was also echoed by many participants.

Documentations of the conference program, abstracts, speaker bios and outcome are available on the conference website (https://events.library.cmu.edu/aidr2019/program/). A subset of slides and posters are published at F1000Research AIDR collection (https://f1000research.com/collections/aidr), and the post-conference proceedings of selected papers are published in Artificial Intelligence for Data Discovery and Reuse 2019 Proceedings [1].

6 ACKNOWLEDGMENTS

This conference is supported by the National Science Foundation Directorate for Computer & Information Science & Engineering (NSF CISE) grant number 1839014. We thank members of the Program Committee, Organizing Committee, and volunteers for your hard work in making AIDR 2019 possible, and thank Carnegie Mellon University Libraries and Pittsburgh Supercomputing Center for administrative support.

REFERENCES

[1] 2019. *AIDR '19: Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse.* Association for Computing Machinery, New York, NY, USA.
[2] National Science Foundation. 2015. Today's Data , Tomorrow's Discoveries. (2015). https://doi.org/10.1108/17260531111151087
[3] Wang. 2019. Artificial Intelligence for Data Discovery and Reuse Conference Website. (2019). https://events.library.cmu.edu/aidr2019/