iConference 2021, March 28, 2021, Beijing, China

Machine Learning and Artificial Intelligence for Science of Science and Computational Discovery: Principles, Applications, and Future Opportunities

Daniel E. Acuna[1], Tong Zeng[2], Han Zhuang[1], Lizhen Liang[1]
(1) School of Information Studies, Syracuse University, Syracuse, NY, USA
(2) School of Information Science, Nanjing University, China

Abstract

Understanding knowledge boundaries, proposing innovative ideas, and producing correct results have become increasingly more competitive in science. Most of these steps rely on colleagues, mentors, and peers. This reliance on humans might not be sustainable because of the growing number of people and ideas entering science. Recent datasets of the scientific enterprise (e.g., full-text publications, citations) offer unprecedented opportunities to solve this issue by using Machine Learning (ML) and Artificial Intelligence (AI). These fields can help the Science of Science (SciSci) and Computational Discovery (CD) understand and automate parts of the scientific process. In this workshop, we propose to 1) introduce participants to principles of modern ML and AI and 2) survey how these techniques are currently used in SciSci and CD. In the end, participants will have a well-rounded understanding of the opportunities and challenges that ML and AI offer.

Purpose and Intended Audience

The Science of Science (SciSci) studies Science itself with the scientific method. It investigates various aspects of the scientific process using quantitative methods to understand the organization, mechanism, evolution, impact, and improvement of scientific activities. Many of SciSci research's guiding ideas could be traced back to the 1930s, taking inspiration from other fields such as Meta-Science, Meta-Knowledge,, and Bibliometrics. The distinctive feature of SciSci is its use of large, heterogeneous datasets about the *doing* of science, including large citation networks, full-text articles, mentorship networks, and success measures (Fortunato et al., 2018; Acuna et al., 2012). Similarly, advancements in computational techniques and datasets about science have allowed researchers to develop methods for Computational Discovery (CD): the partial automatization of processes traditionally done by scientists such as knowledge discovery, evaluation of ideas, and validation of results (Evans and Rzhetsky, 2010; more recently Thsitoyan et al., 2019). Having attended and published in the iConference before, we discovered a gap in previous years' workshops: they lacked the presence of the ever-growing use of machine learning and artificial intelligence in SciSci and CD. With this proposal, we aim to close this gap with a half-day workshop that will teach principles and techniques to a broad set of attendees. We will pay special attention to include historically under-represented disciplinary and demographic audiences. After this workshop finishes, attendees will have a good understanding of SciSci, and CD but will also grasp limitations and opportunities for future research.

We now will explain the purpose and intended audience of the workshop in more detail:

The purpose of this workshop is to:

1. Introduce researchers to the Science of Science (SciSci) and Computational Discovery (CD) research communities
2. Demonstrate and help researchers interested in getting started with Machine Learning and Artificial Intelligence
3. Allow practitioners of SciSci and CD multiple opportunities to interact and network with the organizers, and peers.

This workshop's intended audience is researchers from all research areas in critical information issues that affect contemporary society. These researchers include Information Scientists, Network Scientists, Data Scientists, Computer Scientists, and Librarians. Programming experience is preferred but not required.

Background information. With the development of the Internet, scientific literature has been transformed into digital formats that are indexed, linked, and readily available. Together with other large scale datasets produced by the scientific process, they form the "big scholar data." Recently, there has been an unprecedented release of these digital artifacts for researchers to pursue, including the PubMed Open Access full-text dataset, the Microsoft Academic Graph citation dataset, the Crossref metadata dataset, and the Federal Exporter funding dataset. These datasets offer tremendous opportunities to find relationships between various entities (e.g., funding agencies, institutions, researchers, citizens) and activities (e.g., grant applications, research workforce, publication).

To fully exploit these newly available scientific datasets, we need to use modern Machine Learning (ML) and Artificial Intelligence (AI) techniques to discover, predict, and unfold latent patterns and find and forecast future trends. ML/AI aims at developing algorithms that allow computers to learn from data without being pre-programmed (Stuart and Norvig, 2002). These techniques can be used for learning patterns in text, images, video, and audio. Thus, they are highly suitable for analyzing the large datasets that SciSci uses. They can also help scientists discover new ideas, predict future innovations, and validate results. Interestingly, the ML and AI techniques and applications have remained mostly unknown for a portion of researchers attending the iConference. This workshop aims to help bring awareness to ML and AI partially.

Proposed Format

The proposed workshop will be a combination of presentations by the organizers, hands-on coding experiments with Jupyter notebooks on curated datasets, and brainstorming of ideas about machine learning and its application. We are especially interested in sharing experiences, identifying the challenges, and inspiring future opportunities.

	Time	Activity
B1	9:00 AM - 9:10 AM	Welcoming: goals, format, speakers, and schedule of the workshop
	9:10 AM - 9:40 AM	Introduction to Science of Science: A broad overview of the scale and growth of science vs. scientists, biases, novelty, and problems in peer review, issues of false results, and non-reproducible science.
	9:40 AM - 10:00 AM	Talks about applications of AI and ML to SciSci and CD.
B2	10:00 PM -10:30 PM	Introduction to ML and AI
	10:30 PM - 11:00 PM	Packages and Frameworks: ● Spark: Data transformation, ML pipeline ● PyTorch: Basics, Pretrained models
Coffee break		
B3	11:10 PM - 12:10 PM	Hands on experiments: ● An curated dataset ● Exploratory data analysis ● Modeling with Linear, Tree-based and Deep Learning models
	12:10 PM - 12:30 PM	Flash Talk, 3 or 5 minutes each, the audience share their insights and experiences, could be any topic covers: Interests and benefits; Challenges and Constraints; Opportunities and Future directions

Engagement

We will heavily advertise the website over Twitter, Facebook, and Chinese social media platforms. We will promote the workshop to the list of attendees and interested participants in the Science of Science Summer School (S4), to happen at Syracuse University's iSchool in Summer 2021.

Outcomes

All workshop materials will be shared on https://scienceofscience.org, and the code and discussions will be made available in an openly accessible code repository.

Relevance to the iConference

There are many recent papers about ML, AI, SciSci, and CD published in the iConference (e.g., Hu et al., iConference 2020; Wang et al., iConference 2020; Zeng et al., iConference 2019). However, during the last few years, there has been no workshop covering these topics. We believe this is a lost opportunity as iConference is an ideal place to present the highly interdisciplinary ideas of

SciSci and CD. Given ML and AI's successes in industries and academia, more and more researchers in iSchools are studying and applying ML and AI. We believe that the workshop's focus will attract a good number of interested parties.

Duration: The workshop will be a half-day event.

Attendance: The workshop will host at most 50 attendees.

Special Requirements: No special requirements.

Further reading

- Hu, H., Deng, S., Lu, H., & Wang, D. (2020, March). A Comparative Study on the Classification Performance of Machine Learning Models for Academic Full Texts. In International Conference on Information (pp. 713-737). Springer, Cham.
- Wang, R., Zhang, C., Zhang, Y., & Zhang, J. (2020, March). Extracting Methodological Sentences from Unstructured Abstracts of Academic Articles. In International Conference on Information (pp. 790-798). Springer, Cham.
- Zeng, T., Shema, A., & Acuna, D. E. (2019, March). Dead science: Most resources linked in biomedical articles disappear in eight years. In International Conference on Information (pp. 170-176). Springer, Cham.
- Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Predicting scientific success. *Nature*, *489* (7415), 201-202.
- Evans, James, and Andrey Rzhetsky. "Machine science." Science 329.5990 (2010): 399-400.
- JF Liénard, T Achakulvisut, DE Acuna, SV David, Intellectual synthesis in mentorship determines success in academic careers, Nature communications, 2018
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Vespignani, A. (2018). Science of science. *Science*, *359*(6379).
- Tshitoyan, Vahe, et al. "Unsupervised word embeddings capture latent knowledge from materials science literature." Nature 571.7763 (2019): 95-98.
- Zeng, T., Acuna, D.E. (2020), Modeling citation worthiness by using attention-based Bidirectional Long Short-Term Memory networks and interpretable models, Scientometrics, Scientometrics, 124(1), 399–428.
- Zeng, T., Acuna, DE (2020) Dataset mention extraction in scientific articles using a BiLSTM-CRF model Chapter 11 in Julia I. Lane, Ian Mulvany, and Paco Nathan (Ed.), Rich Search and Discovery for Research Datasets: Building the next generation of scholarly infrastructure, New York