

White Paper on the IMLS Machine Learning Grant

Alex Papson, Anastasia Guimaraes, Christina Leblang, Daniel Johnson, Donald Brower, Eric Lease Morgan, Helen Hockx-Yu, John (Zheng) Wang, Laurie McGowan, Mark Dehmlow, Melissa Harden, Rebecca Leneway

Publication Date

08-12-2023

License

This work is made available under a Exclusive rights in copyrighted work license and should only be used in accordance with that license.

Citation for this work (American Psychological Association 7th edition)

Papson, A., Guimaraes, A., Leblang, C., Johnson, D., Brower, D., Morgan, E. L., Hockx-Yu, H., Wang, J. (Zheng) ., McGowan, L., Dehmlow, M., Harden, M., & Leneway, R. (2020). *White Paper on the IMLS Machine Learning Grant* (Version 1). University of Notre Dame. <https://doi.org/10.7274/r0-320z-kn58>

This work was downloaded from CurateND, the University of Notre Dame's institutional repository.

For more information about this work, to report or an issue, or to preserve and share your original work, please contact the CurateND team for assistance at curate@nd.edu.

**Investigating the National Need for Library Based Topic Modeling
Discovery Systems White Paper**

Table of Contents

Introduction	3
Methodology	4
Bibliography	5
Survey	6
Workshops	7
Findings and Recommendations	10
Findings	10
Recommendations	12
Conclusion	16

Introduction

In August of 2018, the Institute of Museum and Library Services (IMLS) awarded a National Leadership/Planning grant to the University of Notre Dame in the amount of \$49,985 ([LG-72-18-0221-18](#)) to investigate the national need for library based topic modelling tools in support of cross-disciplinary discovery systems. The grant would enable our project team to conduct a series of workshops where we could bring together communities of expertise (computer scientists, librarians, disciplinary scholars) from diverse organizations (large and small universities and colleges, cultural heritage organizations, and governmental organizations) to understand unique current practices of machine learning and to identify possible ways to use topic modeling and natural language processing (NLP) to enhance or augment current library classification in an effort to meet current cross-disciplinary research needs.

Building on our experience developing the cross-disciplinary research tool, Convocate (<https://convocate.nd.edu>), which brings together the disciplines of international human rights law and Catholic teaching, we sought to determine the value of such tools for research, to build a community that could share ideas and experience, and to determine how to advance the use of machine learning in cultural heritage organizations and scholarship. Our underlying thesis was that as universities increasingly focus on cross disciplinary research, libraries have found that their existing discovery tools do not semantically traverse multiple, disparate academic domains. The work to support cross-disciplinary research is itself a diverse intersection of professional concerns — expertise in classification is typically held by librarians, extensive understanding of domain research is held by scholars, and strong competency in computer learning and NLP is typically held by computer scientists. As automation provides more efficiency for traditional library functions, libraries aim to find new ways to provide value for their organizations while also fulfilling their mission to serve patrons in learning, creative inquiry, research, and knowledge/information management. The profession as a whole is transforming from supporting research through scholarly resource acquisition and access to collaborative immersion in the creation of scholarship itself.

The expected output from this project was a white paper (this document) based on our findings. If the results from the community engagement proved positive about the need and interest, the next step would be to organize a diverse working group composed of interested institutions that attended the workshops who would be charged with developing a comprehensive plan for the next phase of the project — to conduct a research program on how to best apply what the team has learned in the support of cross-disciplinary research. As a part of this effort, the cross-institutional committee would apply for a research grant to design an optimized workflow for developing automated metadata and classification for cross-disciplinary discovery.

Since the planning grant was intended to determine the national need, success would be measured through several dimensions as outlined in our application. The project team:

1. would determine the national need for an automated tool that supports cross-disciplinary research;
2. would be able to demonstrate diversity in the representative user groups that it plans to invite. Participants would be made up of scholars, computer scientists, and librarians from academic institutions of varying size and resources. The team would ensure that participants were chosen to represent a variety of use cases and practices in the series of workshops;
3. would identify a likely cohort that could help library professionals pursue further collaborative efforts in improving a cross-disciplinary research agenda;
4. would engage a conversation among scholars, computer scientists, and librarians to reconcile domain-specific, best approaches to supporting cross-disciplinary discovery; and
5. would document current thoughts, models, practices, and tools. This would contribute to the professional literature and hopefully inspire more peers to jump-start similar projects at their institutions. The workshops would be structured to lead participants to raise a rich set of questions and potential solutions.

The artifacts from this grant supported project — workshop presentations, recordings, transcriptions, etc. — are deposited in the [Open Science Framework](#) and the report and appendices are preserved in [CurateND \(the Notre Dame Institutional Repository\)](#).

Methodology

The project team devised a strategy comprised of 3 major activities — 1) create a [bibliography](#) based on an analysis of the scholarly literature to determine what had been investigated already and also to discover who might already be engaged in machine learning activities from the target communities (disciplinary scholars, computer scientists, and librarians/curators) to bring together in a year-long conversation about machine learning and cross-disciplinary discovery, 2) conduct a [survey](#) of higher education and cultural heritage organization stakeholders to better understand the need for expanding machine learning for cross-disciplinary discovery and to surface more potential participants, and 3) to conduct [workshops](#) across the country with a diverse group of stakeholders to build a community of experts that would engage in presentations, collaborative activities, and conversations in an effort to expose the needs, successes, and challenges with utilizing machine learning tools in the pursuit of developing cross-disciplinary discovery.

Towards the end of the initial grant period (December 2019), we determined it could be valuable to bring a select group of workshop participants together once more for a writing workshop that would result in the publication of an edited volume of essays covering various ideas and concerns related to machine learning, scholarly experiences with it, its utility, and its limitations.

The result is [a volume of 14 essays](#) covering an array of machine learning topics that will be published openly at the same time as we submit this report.

Bibliography

As part of the draft of our proposal, our team researched current scholarship related to machine learning in service of cross-disciplinary discovery (see [Appendix A](#)). The goal of this activity was to develop an understanding of the current state of maturity for our thesis across our target communities. Our research turned up 37 articles that we were able to classify into 6 broad categories: classification, cross-disciplinarity, discovery, machine learning, natural language processing, and topic modeling. While our research focus started with exploring how machine learning could support cross-disciplinary discovery, we gave some latitude in our exploration process to find articles that might be on the periphery, e.g. support different aspects of multidisciplinarity, discovery, and/or topic modeling. Relevant topics for readings we discovered included technology in support of metadata creation, methodologies for building vocabulary hierarchies, clustering topics for digital libraries, technology to support transdisciplinary research, ethics and authority control, bibliographic data mining, data modeling in ontology creation, automatic construction of keywords, scholars who teach in cross-disciplinary fields, writer sentiment classification, identifying emerging topics using topic modeling, text mining, cross-disciplinarity and bibliographic/text classification/metadata, automated facet creation, and others.

Certainly, the body of research on machine learning, particularly in computer science, is very broad already, therefore we specifically focused our research on topic modelling in cross-disciplinary discovery. What our research demonstrated is that while there has been some work on the topic of topic modeling in service to cross-disciplinary discovery there hasn't been substantial focus on the topic. It was an indicator to us that there was a need for continued exploration.

Survey

Our first activity to pull together a community of interest was to conduct a survey (see [Appendix B](#)) that could ascertain the experience and interest related to the use of machine learning in cross-disciplinary research for each of the three communities we intended to bring together: librarians, computer scientists, and disciplinary scholars. The responses would not only help us determine topics of conversation for our workshops, but they would also provide some leads for individuals in target communities who would be well suited to participate in our workshops. The questions ranged from demographic information to questions related to individual experiences with machine learning.

Our survey was distributed through various community listservs and our professional networks. It drew approximately 350 (n) respondents with a slightly higher number of librarians and a lower number of computer scientists, but a moderate number of responses from disciplinary scholars.

Librarians: 161 (45%)
Teaching and Research Faculty: 96 (27%)
Computer scientists/engineers: 30 (8%)
Other: 68 (19%)

It should be noted that this survey does not provide statistically significant results based on the number of respondents (low n) and the distribution of the survey (largely self-selected and not randomized). However, it does provide some indication about the reach of machine learning in libraries and scholarship as well as a sense for how mature the utilization of machine learning is in cross disciplinary research related applications.

Our survey found that 55% of computer scientist respondents, 52% of disciplinary scholar respondents, 31% of librarian respondents, and 42% of other respondents had used machine learning in their research. For those who had used machine learning in their research, 40% used supervised learning (classification and identification), 34% used unsupervised learning (topic modeling), 20% used reinforced learning (neural networks), and 6% used other methods. The survey also indicated that 59% of computer scientist respondents, 62% of disciplinary scholar respondents, 54% of librarian respondents, and 70% of other respondents were moderately to extremely familiar with machine learning. Thus far, the results suggested that our survey had found a reasonable number of people who were familiar with various machine learning applications.

The respondents were overwhelmingly positive (90%) about whether machine learning could be used to enhance cross-disciplinary research, and most (84%) indicated that they collaborate with scholars from other disciplines in their research. We also asked respondents to indicate which cross-disciplinary subjects (language and literature, history, fine arts, philosophy/theology, other humanities, engineering, chemistry, physics/astronomy, other sciences, business/economics, law, political science, sociology/psychology, and other social sciences) that each community participated in. Disciplinary scholars had a fairly even distribution related to those they collaborate with — chemistry and physics/astronomy in the 4%

- 5% range, history and language/literature in the 26% - 31% range, and the rest ranging from 11% - 19%. Librarians showed a similar collaboration distribution with chemistry and physics/astronomy in the .6% - 1% range, history and language/literature in the 14% - 16% range, and the rest ranging from 2.5% - 8%. Computer scientists tended to collaborate with humanists (50% in aggregate, 10% fine arts, 13% history, 13% language/literature, 7% philosophy/theology, 7% other humanities). Collaboration in chemistry and other social sciences came in at 0% and the rest between 3% and 7%.

We also asked respondents about their organizational readiness in handling obstacles to cross-disciplinary research. 31% said yes (they were ready), 35% said maybe, and 35% said no (they were not ready). With only 35% in the affirmative category, this was an indicator to us that machine learning was relatively nascent outside of computer science and represented an area for both growth and interest amongst our community.

A few interesting results from the survey included obstacles to conducting cross-disciplinary research. A supermajority of respondents (76%) indicated that terminology/jargon is an obstacle to cross-disciplinary research, a result that suggests it would be beneficial to investigate how machine learning could assist with cross-disciplinary discovery. Other obstacles cited in the survey included funding sources, lack of computing resources, priorities of individual scholars, and finding the right people to support the work. Finally, we found that computer scientists were less likely to collaborate in general, with 52% indicating they typically didn't collaborate outside of their discipline, and we found that overwhelmingly librarians collaborate with librarians. When computer scientists do collaborate, however, they more often collaborate with humanists. This last result indicated to us there was a strong need to bring these three disparate communities together. Diversity in thought is known to produce greater creativity.

Workshops

The heart of our work was to bring our three target communities together in an effort to collaborate, educate, and confer on the topic of machine learning and cross-disciplinary discovery (see [Appendix C](#)). We wanted the workshops to have a diverse representation of computer scientists, librarians, and disciplinary scholars from a range of institutions both in size and in type. With this in mind, we arranged for workshops on the East Coast, West Coast, and in the Midwest to facilitate participation from institutions in different parts of the country and to reduce the burden and cost of participation. Attendees came from all over, representing both large and small Universities (Michigan, Michigan State, Stanford, Yale, Notre Dame, North Texas, Purdue, Indiana, Illinois, Nebraska, DePaul, British Columbia, Northern Arizona, Cincinnati, Georgia, Oklahoma State, San Jose State, Columbia, Boston College, SUNY, Case Western Reserve, Northeastern, Harvard, Edinboro, Rutgers, MIT, CUNY, Rhode Island, Virginia, Utah, Catholic University, George Washington, George Mason, Duke), colleges (Haverford, Berea, Lafayette, Saint Mary's, Pratt Institute), one HBCU (Morgan State), specialty libraries and museums (The Getty, PBS, the Library of Congress, the Digital Public Library of America, the United States Holocaust Memorial Museum, the National Library of Norway, the National Library of Medicine, the Folger Shakespeare Library, the Federal Reserve Board), a few companies (Kyndi, Elsevier, Vantage Solutions), and even one municipality (City of San

Jose). More than 245 people expressed interest in participating in the four workshops, which was more than what the overall budget allowed. 95 individuals were selected to participate in person, and around 10 virtually. Some of the virtual participants stayed for only part of the workshops. We selected for diversity in participants as well, seeking a blend of backgrounds, genders, etc. Our gender distribution ended up being roughly 63% male to 37% female. We would have liked a better distribution of gender, but unfortunately, we didn't get enough female applicants. It was challenging to create a complete balance of all of the categories due in part to the self-selective nature of those expressing interest, but the attendance rosters do reflect how our selection process emphasized building the most diverse group possible from those who applied. In the future, we will want to investigate ways to encourage more female participation.

Every workshop started with a round of presentations to get people thinking about machine learning, to create educational value, and to seed ideas for later discussion. Presenters from a variety of institutions gave conference-quality talks covering a wide range of machine-learning-related topics. The presentations at Notre Dame covered topics such as the importance of human review in the training process to ensure good quality outcomes from machine learning, how machine learning can be used to create a more useful recommender system for library resources, how machine learning could be used to learn more about archival digital image collections, and how artificial intelligence can speed up analytics so people can spend more time on analysis that has to be done manually. The range of topics in Palo Alto examined appraising, processing, and providing access to email archives using a Stanford-developed tool called ePADD, a machine learning platform developed at the University of Cincinnati that supports a range of disciplinary scholarship, another deep learning machine developed at Haverford College that works with handwriting recognition, how Kyndi, a commercial AI software company, developed a platform to help professionals synthesize large sets of information, and lastly a project that is using machine learning to interpret brain waves. The New York workshop featured presentations on many different considerations needed when working with machine learning, a project to analyze undergraduate student writing and how domain experts could formulate questions that can be answered by machine learning, how the Digital Public Library of America is using machine learning to indicate metadata quality to contributors across the country, a project that compared traditional machine learning techniques (support vector machine and linear discriminant analysis) to convolutional neural networks in being able to identify giraffes in photographs, and a means for creating keywords from documents in the Freedom of Information Archive to create a usable form of metadata. And finally, the Washington, D.C. presentations covered topics such as using machine learning tools in combination with human analysis for automated metadata creation for a large scale archive, how the digital scholarship lab supports innovative scholarly projects at Yale, a project using machine learning to determine how government regulations have affected different industries over time, a project to combine usage statistics and reader sentiment analysis to predict circulation patterns, a project to automate creation of metadata using machine learning on images, and the moral and ethical dimensions of machine intelligence.

At the Notre Dame workshop, after the presentations we spent the rest of the morning in a group discussion on pre-arranged topics including tools people used in their projects and deficits in tools for cross-disciplinary research. We found that whole group discussions such as these did not cultivate the level of conversation that we were hoping for. In a large group, people were more hesitant to speak and required quite a bit of coaxing. In an effort to improve the workshops, we changed the second activity to a brainstorming project, in which participants

took 5 to 10 minutes to write down their thoughts on post-it notes related to four categories: 1) common tools used to overcome cross-disciplinary research obstacles, 2) cross-disciplinary research challenges, 3) successful strategies for cross-disciplinary research, and 4) cross-disciplinary problems that can be solved and those that cannot. The ideas were then grouped thematically, and the moderator asked people who had written particular ideas to explain a little about what they were trying to convey. We found that this approach encouraged substantial conversation as people were more inclined to share their thoughts in writing and then comment on them in the broader group setting. The conclusions of these discussions and the afternoon breakouts will be covered in the “Findings” section, and transcribed charts of the participants' responses are posted in [Appendix C](#) with the details related to the workshops.

For the afternoon's activities, we wanted to create in-depth, engaging conversations across our different communities. We organized break out groups to account for diversity in community, gender, and organization type/size. We wanted each group to have a variety of perspectives so that we could encourage cross-community learning. Using a tool called Slido, we let the participants generate topics of discussion related to machine learning and cross-disciplinary research. The participants then voted to determine the most popular discussion topics. Breakout groups discussed topics of interest. The Notre Dame session only had one breakout, but as noted, we found this activity to be engaging for participants, so moving forward, we removed the big, open group discussion from earlier in the day and added a second small group discussion to the rest of the workshops. After the small groups discussed their topics for an hour each, they came back to the larger body and reported out what they discussed. Notes from these reports are shared in [Appendix C](#).

At our first workshop at Notre Dame, the various breakout groups had conversations about library collections as data for use in machine-learning-based research, the ethical considerations around machine learning, machine learning for automated collection metadata and description, and how machine learning technology can aid humanities research even for those who don't understand algorithms. In Palo Alto, breakout groups discussed the degree to which it was important for scholars to understand the mathematical principles built into algorithms, when it is better to train an algorithm from scratch as opposed to using an existing algorithm and refining it for a project, what different machine learning approaches exist and how to find out more about the potential utility for different research goals and data types, what library-centered artificial intelligence could look like, and what infrastructure should academic libraries provide to support machine learning. In New York, breakout groups discussed the ethical considerations of machine learning, library roles in data management for machine learning, the importance of unstructured and serendipitous access to cross-disciplinary content to discover unexpected connections, the degree to which machine learning can be used in libraries, how machine learning can be applied to metadata creation, and how machine learning should be taught in higher education. The Washington D.C. groups discussed what types of people should be on cross-disciplinary teams, how machine learning can be applied to metadata creation, how to conduct name/entity extraction from archival collection descriptions, and how to automatically classify digital collections from their content.

As can be seen from the different breakout groups across the various workshops, there were some topics that were popular at several of the workshops, especially ethics, automated metadata creation and classification, and how libraries can best support machine learning in scholarship.

Because the grant team stewarded our funds well, we were able to add an additional workshop — a writers’ workshop with the goal of publishing an open access edited collection of essays. The authors’ planning session took place on October 25, 2019 at the University of Notre Dame. The workshop was designed for participants to share ideas, work with colleagues to refine their hypotheses and arguments, and to construct an outline for their essays so they could begin writing upon returning from the workshop. The writers’ workshop attracted 14 on-site attendees and 2 remote participants who contributed 14 essays. The results are published as [*Machine Learning, Libraries, and Cross-Disciplinary Research: Possibilities and Provocations*](#).

Findings and Recommendations

Findings

Once the workshops were complete, team members reviewed the workshop recordings, collocated results from the different activities, and summarized the discussions, presentations, and reports. This section on findings and recommendations is an amalgamation of the various workshop activities and is organized in broad concepts. For more detailed information, please see [Appendix C](#).

1. Interest in Machine Learning is High and Appears to be on a Precipice

We had no difficulty identifying people to participate in our survey. Workshop attendees were enthusiastic and passionate about the topic, and there are an increasing number of professional meetings on the topic of machine learning. Many people are interested, and the interest seems akin to the interest in open source software twenty-five years ago.

2. The Biggest Issues with Cross-Disciplinary Research are not Discovery Related

We defined “cross-disciplinary research” as a scholarly endeavor including diverse subject matters such as physics and theology, economics and engineering, or musicology and agriculture. The diversity of norms of scholarship between such disciplines such as the modes of scholarly communication, the roles of quantitative and qualitative analysis, or the degree of interpersonal collaboration are not addressable by something like machine learning. Machine learning is a tool for prediction, classification, and clustering, it is not a tool for addressing norms of behavior.

3. There is a High Need for Interdisciplinary Collaboration

To put machine learning into practice requires three distinct sets of knowledge: 1) domain expertise, 2) statistical expertise, and 3) computer programming expertise. None of these knowledge sets are trivial to obtain and each requires years of experience to completely understand. Thus, it is unrealistic to believe any one person can do machine learning without the assistance of others. As a corollary, if a discipline is not amenable to collaboration, then the discipline will be less amenable to the use of machine learning

4. Community Effort for Greater ROI

As canned algorithms and many open-source algorithms may still be advantageous to libraries and scholarship, it is evident from what we learned that academic communities need to tinker with them to enable them to work with library data effectively.

5. “Garbage in, Garbage out,” Machine Learning Requires Good Data

The process of machine learning turns decision-making on its head. Instead of writing sets of rules used to make a decision, sets of observations (information) are given to a computer program to make a prediction. The observations are converted into “vectors”, aggregated, and saved as a “model.” New information is then compared to the model to determine how something might be classified, clustered (grouped), or ordered. This process has been used to classify novels according to genre or suggest what item a person might purchase next. While machine learning often works, it is only as good as the data given to it. If the data includes false or biased information, then the results will be false or biased. The output of machine learning is only as good as the input.

6. Ethics are a Really Big Concern for Machine Learning, Especially Regarding Bias

Machine learning can be a very powerful tool, and it has already been woven into our lives. For example, it is used to predict the weather with an uncanny degree of success. It is used to help us get from Point A to Point B quickly and easily. It helps us write and search the Internet through the use of auto-complete functions. Such powerful tools can be used with good or bad intent and require knowledge of individual behavior that could compromise a person’s privacy. An implementation of machine learning often requires significant amounts of information which, in turn, makes the results sometimes seem like magic. The ethical use of machine learning has yet to be fully articulated.

Data bias associated with canned algorithms is one of the biggest challenges facing the

use of machine learning. Although Amazon and Google trained their machine learning algorithms with an exceptional amount of data, we learned that the benefits gleaned by many other industries could not directly carry over to libraries. The main reason is that those commercial entities hold a different type of data than libraries. For example, the above companies trained their image processing and facial recognition algorithms with “modern” data. However, many libraries hold historical collections, data that do not exhibit similar “patterns” or “characteristics” that machines derive from modern images. One of the fascinating cases that the University of Utah identified was that an algorithm mistakenly categorized a letter that people were reading as a laptop. In many cases, the current environment and our cultural, social, and technological contexts have shifted dramatically from that of library collections and the canned algorithms produce inaccuracies.

7. There is a Need for Greater Machine Learning Literacy

When it comes to artificial intelligence and machine learning embedded in our lives it is important to understand what machine learning can and cannot do. A new form of information literacy is needed for machine learning. We need to learn when to trust the output of machine learning, when to take it with a grain of salt, and when to ignore it altogether.

Recommendations

From our workshops and the survey, it is obvious that machine learning for cultural heritage, library, and scholarly use outside of strictly computer science disciplines is at a relatively early stage. The survey respondents and workshop participants expressed challenges and obstacles to either start, engage, or further their efforts in exploring the potential of machine learning for operations, scholarship, and services. Some of the respondents did not feel that they have the expertise to grasp even where to begin while some have been experimenting but have limited capacity and bandwidth to focus and do more. Institutionally speaking, libraries are lacking a general commitment to explore machine learning, due to enduring responsibilities in their current service offerings. The grant team heard diverse needs from our target communities, and the report provides the following set of recommendations based on what we heard. Our team believes that these recommendations could assist the library community in progressively and strategically approaching machine learning.

1. Increase the the Community

There is broad interest in machine learning, and this interest resides locally, regionally, nationally, and internationally. Here at Notre Dame we discovered siloed machine learning projects. The Hesburgh Libraries, as a rather centralized resource, could facilitate communication between leaders of these projects with the intent of sharing knowledge and thus increasing understanding across the University. Through our local workshops, we established relationships with other academics in nearby universities. These relationships are ripe for collaboration. We might reach out to learn others' desire in this same regard. Increasingly, there are library-based national conferences with machine learning and artificial intelligence tracks. We plan to participate in these conferences to share our experiences and to continue building community. There is also a fledgling international venue for libraries and machine learning — Fantastic Futures. In 2018 it was held at the National Library of Norway, and in 2019 it was held at Stanford University. Next year it is planned to take place in Paris. We have participated in both meetings, and we hope to participate again this year. Finally, we believe downstream it might be possible to create some sort of membership organization that would have resources to spend on both community building and resource sharing, but we also believe such an organization is a bit premature at this stage.

2. Develop Machine Learning Education for Scholars and Library Professionals

Many participants expressed the necessity for learning. Learning needs reflect where the field is currently, with special interests in tools, algorithms, mechanics, training, and data bias. Education is also crucial for adopting machine learning into specific contexts. For example, while a variety of general or industry-specific machine learning courses and lessons are available online, few are tailored for library work, such as collection processing, metadata creation, and knowledge discovery.

A facilitated workshop might include topics such as: 1) what is machine learning and why should I care, 2) what are the differences between types of machine learning (classification, clustering, regression, and dimension reduction), 3) collecting and preparing data for machine learning (a process called, "vectorization"), 4) selecting algorithms, creating models, and validating results, 5) applying models to solve real-world, operational problems, 6) discussion and summary. These workshops would require little to no programming experience, but they would require a willingness to do work from the command line interface. These workshops could be facilitated locally, regionally, nationally, internationally, or even virtually to groups of no more than thirty participants at a time. In the end, participants would have a greater understanding of what machine learning can and cannot do.

3. Form Learning Communities and Networks

Our investigation revealed that there is a sizable machine learning early adopters group forming. It is paramount to connect them with other peer institutions, which will foster the exchange of ideas and inspire supported learning and innovation over time.

Organizations may consider seeking a startup grant that creates a conference solely devoted to machine learning. The fund may bridge conference operations for the first three to five years, and eventually, the event can be self-sustained through conference registration fees and sponsorship. The project team heard the desire from participants for such an opportunity to build a machine learning community of practice and professional network. The team also recommends that the event continue to cultivate relationships among scholars, librarians, and software engineers. Rather than envisioning an event geared strictly to librarians, it could be a conference focusing on the creation of practical solutions via collaborations among all three communities.

4. Create and Curate a Clearinghouse for Machine Learning Models

Machine learning models may represent a new component of library collections. Machine learning is a two-step process — the first step is the creation of a model, and the second step is the model's application. Like a book, a lot of intellectual effort goes into the creation of the model, and once it is created, it can be used by many people for their own purposes. If a group of stewards were to collect and curate machine learning models, then the greater community might benefit from degrees of scale. For example, models (which are merely software) could be first described, classified, and cataloged in a centralized repository. The creation of a clearinghouse/collection is fraught with difficulties, but considering the need for the transparent distribution of models complete with credentialing, such as peer-review, creating and curating machine learning models would create significant value.

5. Support Consortia Around Subject Strength to Develop Machine Learning Tools

Studies indicate that algorithms that solve similar problems are comparable in terms of design. The difference in performance depends more on the volumes and quality of data with which they are trained. In the end, whoever owns most of the quality data holds a substantial competitive advantage in machine learning over those who do not. To realize the potential of machine learning, libraries should pool similar data together for machine learning research and development. For example, civil war archives may collaborate to contribute to a "training data set" of civil war images. If the training set covers a variety of characteristics of photographs, it will help any machine learning algorithm to increase accuracy. Libraries have already been forming communities based on collection strength. Pooling the collections from those communities will create

opportunities to develop machine learning tools that increase collection processing and eventually, collection usage. Acknowledging the difficulties to jump-start such peer collaboration, funding agencies and foundations may consider start-up funds to support necessary personnel, expertise, and infrastructure to help those types of endeavors.

6. Develop Processes to Enhance Discovery Tools

Part of our original hope was to use what we learned from this project to evaluate how cross disciplinary discovery could be improved through automation of subject terminology using machine learning. We believe this goal is still valid and could be one of several enhancements that machine learning could bring to discovery tools. Additionally, one of the most common tasks people in libraries wanted to accomplish through machine learning was the classification/tagging of images. The tags would then be used to augment descriptions indexed by our discovery systems. All of this is in the hopes of decreasing library backlogs and increasing access to library collections. While this is a laudable task, it does not maximally exploit machine learning. Increasingly, library collections are born digitally. This means it is possible to have a computer ingest content, analyze for characteristics like parts-of-speech and named-entities, and then augment bibliographic description in an innovative way. For example, it would be possible to count and tabulate all the occurrences of all personal names in a given work. If the count and tabulation was above a pre-articulated threshold, then the work would be classified/tagged with the name. The same thing could be done for places and dates. Similarly, using any number of different algorithms, it is possible to compute statistically significant keywords which could be used to supplement traditional subject headings. The use of the discovery system itself could be used to improve ease of use by making suggestions for further exploration based on usage.

7. Support Diversified Machine Learning Innovations.

As demonstrated by various applications presented at the workshops, institutions exhibited their own unique research interests in machine learning. Although this planning grant goal was for the betterment of the discovery of multidisciplinary materials leveraging machine learning techniques, the team learned that participants were conducting machine learning research in a breadth of areas, aiming to solve their pressing challenges. The team also saw the potential for participants' work to help peer institutions.

In collaboration, funders and the machine learning community may consider sponsoring the diversified development of machine learning applications that respond to local research needs. The funds may also encourage collaboration among scholars, researchers, engineers, and librarians to include all necessary expertise and promote

multidisciplinary research. This diversified machine learning portfolio will enable developing machine learning applications that are responsive to multiple academic needs simultaneously. Any innovation that shows potential could later support forming a partnership for those who face similar challenges.

Conclusion

Returning to the five measures of success outlined at the beginning, we believe our findings will be valuable to the scholarly, library, and computer science communities. We set out first to explore if there was a national need for machine learning to support cross-disciplinary discovery. What we discovered is that there is considerable interest in how machine learning can support scholarship, but efforts to use machine learning in libraries and disciplinary scholarship is still fairly nascent. We had hoped, following the completion of this project, to form a working group that could collaboratively set a trajectory for how to implement machine learning technologies in support of cross-disciplinary discovery. We believe this will be a goal that will have greater traction in the future, but based on the outcomes from the workshops, we believe there is a need first to build a community foundation for machine learning in scholarship before we are likely to get to the more specialized questions related to machine learning and its efficacy in scholarly and library endeavors. While our workshops didn't specifically determine a pathway forward for machine learning in support of cross-disciplinary discovery, the findings did give us some ideas on how to start working towards that direction and we did succeed in getting a diverse group of engaged computer scientists, librarians, and disciplinary scholars together in a series of energized discussions about the present state of the field and in beginning to brainstorm a compelling future together. We suspect the utility of machine learning in the scholarly environment is picking up momentum and will only expand over the coming years.

Our second goal was to demonstrate diversity in participation. Selection for participation was based on a pool of interested individuals and experts we could find through our research. To help mitigate concern about a limited, self-selected list of volunteers, our criteria for participation considered the different professional backgrounds of participants, sizes and types of organization, and a spectrum of experiences with machine learning. While we would have liked to achieve greater gender parity, our workshops did garner participation from large and small colleges and universities from all over the country (including one HCBU) as well as specialty libraries, museums, companies, and one municipality. The variety of backgrounds and institutions strengthened participant engagement and created an event where all had an opportunity to learn something new.

Our third goal was to establish one or more cohorts of expertise and interest that could help spawn further collaborative efforts in the future. The very design of the workshops helped to establish a diverse cohort of professionals who shared an interest in machine learning to support research and content management. As the final part of the project, we brought together over a dozen of the participants to create an edited volume of essays about ideas, issues,

dangers, and benefits of machine learning from the contexts of scholars, librarians, and computer scientists. Given that the broadening of fields beyond computer science is still in its early stages, we hope that this collected work will provide engaging and timely information to the community. We deliberately chose to provide this volume in an open access format to provide benefit to as many people as possible and for publication to happen while the ideas are fresh and relevant.

Our fourth goal was to engage a conversation amongst participants with different backgrounds and expertise to reconcile domain-specific, best approaches to supporting cross-disciplinary discovery, and our fifth goal was to document current thoughts, models, practices, and tools. The artifacts from the workshops and the analysis of the outputs clearly indicate that we had a high level of participant engagement and everyone had something valuable to offer. The workshop presentations and discussions highlighted a wide set of tools, models, and practices as well as participants' thoughts on the current state of machine learning in service to scholarship and cultural heritage organizations. While our workshops were designed to consider machine learning in the context of topic modelling in support of cross-disciplinary discovery, we gave some latitude to explore concerns and ideas more broadly across both machine learning and cross-disciplinary research. This latitude is what helped reveal that the core concern for most scholars and organizations related to machine learning and interdisciplinary research wasn't primarily focused on discovery. Based on our research, survey, and workshops, the most prevalent issues with cross-disciplinary research at this time appear to be cultural and logistical. More cross-community collaboration is needed to be effective in the application of machine learning, more community sharing of algorithms and models could boost others' work, good machine learning requires good data, ethics and bias are the biggest concerns in the direct application of machine learning, and there is a need for broader understanding in the scholarly and library communities about machine learning and how it works.

To this end, we have recommended continuing to build scholarly and cultural engagement in machine learning, and by extension, to create communities of practice and learning networks, develop a basic curriculum to provide broad machine learning training for scholars and other professionals, create a warehouse where learning models can be shared openly, thereby pooling quality data together to improve automating creation of metadata, develop shareable processes to use machine learning to enhance resource discovery, and encourage and support more machine learning innovations. Several of these recommendations could be additional opportunities for the IMLS and other granting organizations to support — community development could be propelled by developing an annual conference on machine learning and its applications in our communities. A course similar to [the Carpentries workshops](#), for example, could be developed to help spread knowledge and competency in machine learning applications, and any number of proposed innovations leveraging machine learning could be supported. We are particularly intrigued by the ideas of continuing the development of the community and building a curriculum to share and extend knowledge of machine learning.

Once this more foundational work is accomplished, we believe the time will be right for advancing the development of cross-disciplinary discovery tools.

We hope these findings and recommendations will be valuable to the cultural heritage and scholarly communities and that they can help to establish a meaningful agenda to coalesce a diverse community of experts interested in machine learning applications in the academy. Machine learning applications in libraries and non-computer-science-related disciplines are still relatively young, and we believe that with the right leadership and cultivation, machine learning stands to have a substantial impact in both operational and scholarly arenas. Within a few years, as our professions gain experience and continue exploring the usage of machine learning, we believe that there will be greater interest in more narrow topics such as topic modelling in support of cross-disciplinary research. Until then, there are many opportunities for our communities to advance the utilization of machine learning. We hope the Institute for Museum and Library Services finds the results of the project as valuable as we did and want to thank them for their investment and support during the grant period. Our discoveries wouldn't have been possible without it.

Submitted, December 2020

*IMLS Investigating the National Need for Library Based Topic Modeling Discovery Systems
Project Team, Hesburgh Libraries, University of Notre Dame*

John (Zheng) Wang, Associate University Librarian, Digital Access, Resources, and Information Technology (PI)

Donald Brower, Digital Projects Lead, Hesburgh Libraries

Mark Dehmlow, Director, Library Information Technology Program

Nastia Guimaraes, Project Management Librarian

Melissa Harden, Web Strategy Librarian

Helen Hockx-Yu, Manager Digital Asset Strategy, Office of Information Technologies

Daniel Johnson, English Literature and Digital Humanities Librarian

Christina Leblang, Associate Program Director, Notre Dame Office of Life and Human Dignity

Rebecca Leneway, Project Manager

Laurie McGowan, Digital Project Manager

Eric Lease Morgan, Digital Initiatives Librarian

Alex Papson, Metadata Librarian