Machine Learning + Libraries

A Report on the State of the Field

Ryan Cordell
Associate Professor
Northeastern University English Department
r.cordell@northeastern.edu

Commissioned by LC Labs
Library of Congress

Version Published: July 14, 2020

Table of Contents

Foreword

Kate Zwaard, Director of Digital Strategy, Library of Congress

In 2019, the Library of Congress embarked upon a series of initiatives to help realize the vision of our first ever Digital Strategy.[1] To "Invest in our Future," we launched "The Season of Machine Learning," led by the LC Labs team in the Library's Digital Strategy Directorate. We wanted to see where this technology might offer opportunities for the Library, help us explore challenges inherent in its application, and provide directions for further research.

LC Labs has been helping the Library experiment and explore emerging technologies since 2017. We have continued to improve our practice to ensure it's both effective and aligned with the values of our institution by working from evidence, being transparent and collaborative, and considering the broader social effects of our technology choices. Machine learning, in particular, has the potential to transform aspects of librarianship, but contemplating its application also requires us to carefully consider the ethical, operational, and social considerations and guardrails necessary. To that end, The Season of Machine Learning included both practical experimentation and contributions to a broader intellectual framework, informed by scholarship, the norms of our professions, and other contributions.

We began by posting an open solicitation for organizations to help us conduct practical experiments. That, in itself, was a learning experience, as we evaluated proposals from a variety of vendors with their own perspectives and values. We awarded the contract to the University of Nebraska-Lincoln's Project Aida team.[2] The collaboration with UNL was designed to provide us evidence in context and to help us and our professional communities make more informed decisions around when and if to apply ML techniques to library and archival collections.

Later in the year, LC Labs convened practitioners for the Machine Learning + Libraries Summit,[3] to help explore concrete best practices for collaboration, success criteria, and steps forward. The summit coincided with our selection of Benjamin Lee's machine-learning-focused project Newspaper Navigator as one of the 2020 Innovator in Residence projects.[4]

To dig deeper into the opportunities and risks presented by machine learning for institutions like ours, we commissioned Ryan Cordell – Associate Professor of English at Northeastern University – to write this report on the "state of the field in machine learning and libraries." The goal was to provide a wide-ranging view into the current applications and practices of applying machine learning in libraries and other cultural heritage organizations.

While we engaged in the process of experimenting with machine learning from the perspective of our own work and context, we hope that this report, and the related information that has surfaced, will be helpful to others as well. As stated in our digital strategy, we aim to "Drive Momentum in

[1] Read the full Digital Strategy at https://www.loc.gov/digital-strategy/.

[2] Explore our collaboration with the Project AIDA team, including their project code and Final Project Report and Recommendations, at https://labs.loc.gov/work/experiments/exploring-ml/.

[3] A summary of the Machine Learning + Libraries Summit: https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

[4] Newspaper Navigator experiment: https://labs.loc.gov/work/experiments/newspaper-navigator/.

our Communities," and we try to work "out loud" with the aim that it provides some value to you and your organizations.

The Library's vision is that "all Americans are connected to the Library of Congress," and we remain wholeheartedly committed to the core values of access, discovery, and meaningful engagement that connect the Library to the largest possible variety of users. My team is incredibly grateful to Ryan Cordell for this report, which is the product of his depth of experience, carefully considered approach, wide and deep perspective, and a lot of hard work. We also owe a deep debt of gratitude to the scholars and practitioners, cited and uncited in this report, whose hard-earned wisdom we benefit from.

We hope you will find this report as informative as we have. Dr. Cordell has offered many thoughtful recommendations for us and the wider cultural heritage community to consider as we continue learning about and experimenting with emerging technologies. If you're interested in learning more about our work, which includes experiments with machine learning and other technologies in libraries and cultural heritage organizations, you can read more on our experiments web page at https://labs.loc.gov/work/experiments. *The Signal Blog*, found at https://blogs.loc.gov/thesignal, is where we and the Digital Content Management Team write about a variety of digital projects at the Library. And to stay up to date with all of our LC Labs work, sign up for our monthly newsletter.[5]

Finally—we would love to hear from you! You can reach us with comments, questions, or other thoughts about this, or anything else we're working on, anytime via email at LC-Labs@loc.gov.

[5]Sign up for the LC Labs monthly newsletter at https://updates.loc.gov/accounts/USLOC/subscriber/new?topic_id=USLOC_182

1. Introduction

The majority of machine learning (ML) experiments in libraries stem from a simple reality: human time, attention, and labor will always be severely limited in proportion to the enormous collections we might wish to describe and catalog. ML methods are proposed as tools for enriching collections, making them more useable for scholars, students, and the general public. ML is posited as an aide to discoverability and serendipity amidst informational abundance. We might imagine, for example, patrons browsing automatically-derived topics of interest across a digital library comprising thousands or millions of texts—more texts, certainly, than typical constraints on labor or expertise would allow us to imagine labelling manually.

We must also acknowledge, however, that neither computers nor digitization created the challenge of scale in library collections. Our current digitized collections, while certainly large, comprise only a small subset of the analog collections held by libraries and other cultural heritage institutions. Acknowledging that scale is not a problem created by digitization reminds us too that libraries have long been central to conversations about how to provide access to information at scale. Indices, catalogs, bibliographies, and subject guides—whether analog or digital—are all facets of a long tradition of library expertise in this area.

Current cultural attention to ML may make it seem necessary for libraries to implement ML quickly. However, it is more important for libraries to implement ML through their existing commitments to responsibility and care. Section 2 of this report will unpack the particular risks ML poses for library work. Here, however, I simply note that events of the past few years have starkly illustrated the ethical, political, and societal dangers of the Silicon Valley ideology encapsulated by Facebook's now-retired motto, "move fast and break things."[6] While that motto may have emerged to advocate experimentation and iteration with software, we have amply seen how broken software can also break people's bodies, livelihoods, relationships, and agency. This report emerges even as Facebook's founder, Mark Zuckerberg, resists calls by civil rights leaders, and some of his own employees, to moderate inflammatory posts calling for violence against protesters advocating for racial justice.[7] At the same time, information studies scholar Safiya Umoja Noble—author of *Algorithms of Oppression*, discussed further in Section 2—reports that algorithmic filters on Instagram—a Facebook company—are wrongly flagging her posts about `#BlackLivesMatter` as false information.[8] In addition, the influx of ML generated text, audio, images, and "deep fake" videos will necessitate enormous public literacy work in the coming years focused on algorithmically-generated material online. By centering ethics, transparency, diversity, privacy and inclusion, libraries can take a leadership role in one of the central cultural debates of the twenty-first century.

In our interview, Kate Zwaard, Director of Digital Strategy at the Library of Congress, emphasized

[6] Hemant Taneja, "The Era of 'Move Fast and Break Things' Is Over," *Harvard Business Review*, January 22, 2019, https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over.

[7] Cat Zakrzewski, "Mark Zuckerberg Spoke with Civil Rights Leaders About Trump's Posts. It Didn't Go Well." *Washington Post: PowerPost Analysis Analysis Interpretation of the News Based on Evidence, Including Data, as Well as Anticipating How Events Might Unfold Based on Past Events*, accessed June 3, 2020, https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/02/the-technology-202-mark-zuckerberg-spoke-with-civil-rights-leaders-about-trump-s-posts-it-didn-t-go-well/5ed55da4602ff12947e81457/.

[8] Safiya Umoja Noble, "Safiya Umoja Noble PhD on Twitter," Twitter, June 2, 2020, https://twitter.com/safiyanoble/status/1267978038072889344.

"how different libraries are from Silicon Valley in the best ways. Libraries can be deliberate about technological adoption" and make decisions based on what is shown to work to the explicit benefit of patrons, rather than shifting wildly with each new technological trend. Zwaard noted that the Library of Congress has been criticized for its slow move to begin acquiring born-digital materials, for example, but argued this slow speed meant some of the technical debates had time to play out and standards begin to emerge, meaning their acquisitions are more likely to be sustainable. "Through the slow and careful adoption of tech," Zwaard emphasized, "the library can be a leader."[9] The LoC has worked to exemplify such leadership in library ML. Through their programming during the "Summer of Machine Learning"; sponsoring reports such as IDA (described in 1.6.2) and this one; fostering of ML work through the Innovator-in-Residence program; and cultivating researcher projects such as Newspaper Navigator, which this report will cite frequently as a model; the LC Labs team at the Library of Congress seeks to model sustainable, interdisciplinary ML scholarship, education, and outreach.

As institutions of memory and community, libraries cannot be bound to destructive ideologies of technological implementation, but must instead model alternative engagements with ML focused on building rather than breaking. Libraries can become ideal sites for cultivating responsible and responsive ML, as that term describes a constellation of technologies that explain data within the contexts of its collection, aggregation, and association. As Ted Underwood has written, "the fuzzy, context-specific models produced by machine learning have a lot in common with the family resemblances historians glimpse in culture." For Underwood, the subjectivity of ML models is precisely where we find their potential for explaining culture, including its biases and exclusions. Underwood sees enormous possibility in research that weds the contextual knowledge of the humanities with statistical methods such as ML: "It is possible to build real knowledge by comparing perspectives from different social contexts. Historians have long known how."[10] If ML builds knowledge by comparing perspectives, then the library's expertise will be required to make sense of this technology and its findings.

Similarly, in an interview sociologist Laura Nelson described ML as a "radically inductive method" rather than a deductive method. As such, ML is a computational domain that requires the epistemologies understood by qualitative scholars, and ML cannot be shoehorned into supposedly objective scientific frameworks. Nelson argues that the contributions of humanists and social scientists should be among the most valuable to the future of ML, but that future will require us to find ways of speaking meaningfully across disciplines.[11] These sentiments echo those of the "Algorithmic Equity Toolkit" team, who argue strongly that while data science "conventionally valorizes engineering techniques to a greater degree than understanding the social domain in which data arise," this approach is short-sighted. Integrating qualitative and participatory scholars and methods, they write, "deepen[ed] the qualitative and reflective dimensions of this work and support[ed] us in enlarging the scope of who should be considered essential to the practice of data science."[12]

[9] Ryan Cordell, "Machine Learning and Libraries Interview with Kate Zwaard," May 27, 2020.

[10] Ted Underwood, "Why an Age of Machine Learning Needs the Humanities," Public Books, December 5, 2018, https://www.publicbooks.org/why-an-age-of-machine-learning-needs-the-humanities/.

[11] Ryan Cordell, "Machine Learning and Libraries Interview with Laura Nelson," March 17, 2020.

[12] Michael Katell et al., "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona Spain: ACM, 2020), 45–55, https://doi.org/10.1145/3351095.3372874.

The need for ML work that stretches across disciplines pertains not only to academic settings. In an article describing how his company built "an AI-based decision support tool for financial analysts" Joseph Byrum notes that diversity of background and expertise is necessary:

> How does one build a team to tackle textual analytics? Parsing meaning is a fundamentally human exercise involving statistics and math, yes, but also linguistics, cognitive psychology, sociology, ethnography, and STEM (science, technology, engineering, and math). It involves multiple disciplines in collaboration, rather than one or two disciplines in isolation.[13]

I cite this example because, in building genuinely interdisciplinary and diverse ML teams—not just in terms of discipline, importantly, but also in terms of the identities of team members—libraries can model effective, just practice for governmental and even corporate sectors. This will require deliberate and sustained efforts to build diverse ML teams, especially given that, on average, neither library nor academic staff currently reflect the diversity of the communities they could serve.

Nelson identified one of the greatest needs in ML today as "translators," mediators who can translate between computer scientists, humanists, and social scientists.[14] These translators must be able to imagine how domain-specific questions could be operationalized for specific ML tasks, which requires interdisciplinary perspective. While this kind of collaboration has happened in the digital humanities and adjacent fields, even there it remains the exception, rather than the rule, that computer scientists and humanists would directly collaborate around a dataset of common interest.

Nelson's call resonates with a recent paper by Eun Seo Jo and Timnit Gebru calling on ML researchers "to take lessons from other disciplines…that have longer histories of addressing" the ethical concerns raised my ML data and methods. "As disciplines primarily concerned with documentation collection and information categorization," Jo and Gebru argue, "archival studies have come across many of the issues related to consent, privacy, power imbalance, and representation among other concerns that the ML community is now starting to discuss." Jo and Gebru advocate for an entirely new, interdisciplinary branch of ML research that blends ML and archival studies:

> By showing the rigor applied to various aspects of the data collection and annotation process in archives, an industry of its own, we hope to convince the ML community that an interdisciplinary subfield should be formed focused on data gathering, sharing, annotation, ethics monitoring, and record-keeping processes.[15]

In 5.5 I discuss means for developing ML expertise among library staff, but Jo and Gebru's argument reminds us that libraries' existing expertise in responsible collection and description already has much to offer broader ML conversations.

Building on their existing roles as hubs of intellectual exchange, libraries could become focal sites for the translation and collaboration that will be required to cultivate responsible ML research. Librarians whose expertise requires consultation and collaboration across domains are well posi-

[13]Joseph Byrum, "Build a Diverse Team to Solve the AI Riddle," MIT Sloan Management Review, May 18, 2020, https://sloanreview.mit.edu/article/build-a-diverse-team-to-solve-the-ai-riddle/.

[14]Cordell, "Machine Learning and Libraries Interview with Laura Nelson."

[15]Eun Seo Jo and Timnit Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning," December 22, 2019, https://doi.org/10.1145/3351095.3372829, pg. 2.

tioned to make connections those working in academic disciplines, technical positions, or corporate domains might miss. From the perspective of project organization and funding, libraries are ideal sites for work that brings together expertise from across campus or across communities. This report proposes that libraries are not simply poised to benefit from practical applications of ML, but instead that libraries are in a position to take leadership in broader societal conversations about responsible ML.

1.1. What is Machine Learning?

The phrase "machine learning" comprises a range of methods through which computers learn from data without being explicitly programmed to generate a particular output.[16] ML algorithms use statistical models to identify patterns in data and can roughly be divided between "supervised" and "unsupervised" methods. Supervised ML approaches benefit more explicitly from human expertise and input, while unsupervised approaches potentially expand our perceptions by identifying new connections in data. Between the poles of supervised and unsupervised one also finds a spectrum of semi-supervised approaches to ML that seek to join human expertise to computational serendipity.

Supervised ML relies on labeled training data to make decisions about unlabeled data. Benjamin Lee writes one of the most concise outlines of a supervised ML workflow:

> The canonical workflow with supervised learning is as follows. First, a set of labeled data is partitioned into a training set, validation set, and test set. A machine learning classifier uses the training set, including classification labels, to learn a classification function. The classifier is then fed the validation set without the corresponding ground-truth class labels and predicts the class labels according to its learned classification function. Based on the performance of the classifier, which is assessed by comparing the predicted labels to the ground-truth labels, the classifier's hyperparameters are tuned, and the classifier is subsequently re-trained with the training set, after which the performance on the validation set is re-assessed.[17]

For example, an ML program that seeks to sort literature into genres might begin with a set of literary texts for which genre has been tagged by domain experts. Using that training data, the program learns the features of each genre, which it can use to sort untagged literature into the existing generic categories.

By contrast, unsupervised ML infers the structure of an unlabeled dataset by finding hidden patterns within it. For example, an unsupervised literature classification program would attempt to group a set of provided texts by internal similarities alone—perhaps based on vocabulary, or sentence structure, or some combination of features—and might produce groups of texts human readers would not readily recognize as distinct genres. Importantly, even unsupervised ML relies on human

[16] For a brief but thorough introduction to machine learning, see Alpaydin, Ethem. *Machine Learning*. The MIT Press Essential Knowledge Series. The MIT Press, 2016. https://mitpress.mit.edu/books/machine-learning.

[17] Benjamin Charles Germain Lee, "Machine Learning, Template Matching, and the International Tracing Service Digital Archive: Automating the Retrieval of Death Certificate Reference Cards from 40 Million Document Scans," *Digital Scholarship in the Humanities* 34, no. 3 (September 1, 2019): 513–35, https://doi.org/10.1093/llc/fqy063, pg. 523.

judgement about what precisely comprises the set of training data. An unsupervised ML algorithm clustering all the books in the HathiTrust digital library would likely cluster quite distinctly from a similar process clustering only books identified as fiction by human readers. The construction of datasets always requires human judgment.

While we sometimes think of computational processes as more neutral or objective than human decision-making, Cathy O'Neil more accurately defines an algorithm as an "opinion embedded in mathematics." Models, O'Neil shows throughout her book, "despite their reputation for impartiality, reflect goals and ideology," while even the question of "[w]hether a model works is also a matter of opinion."[18] In other words, a computational process is "automatic" only insofar as it acts autonomously on rules written for it by human programmers, based on their assumptions about a given domain, and using data gathered and provided by human beings, who made decisions about what to record and what not to record. Section 2 of this report samples more fully from current scholarship about algorithmic injustice, which must be considered as a prelude to conversations about ML in libraries.

Like other computational processes, ML is neither neutral or objective, though its subjectivity can be harder to pin down than that of other kinds of algorithms. Because many ML algorithms learn from the structure of the data they are provided, rather than following an explicit set of instructions provided by a programmer, the opinions embedded in them can seem more obscure, more difficult to parse and assign. When discussing ML, the ethical imperative shifts to the source data. The assumptions, exclusions, and biases of any given dataset will reverberate throughout any ML processes using that dataset. Importantly, however, we cannot simply excuse biases ML processes by decrying the data—the data and ML algorithms are parts of a larger system that must be addressed at all levels. Because of the ways ML forces researchers to confront the structure of datasets, however, it proves an especially fruitful locus for conversations between computer scientists, technologists, and data scientists on the one hand and librarians, scholars, and cultural heritage professionals on the other.

Despite headlines proclaiming "The Librarians of the Future Will Be AI Archivists"—in a *Popular Mechanics* article describing Newspaper Navigator, a project this report will repeatedly praise—the reality will be far more measured. The future will require integrating the expertise of human librarians with ML.[19] As this report will note repeatedly, even practices of collection and annotation that rely entirely on human expertise nonetheless require vigilant evaluation to manage bias and promote justice. Responsible ML in libraries can work in concert with human expertise, bringing attention to oversights and omissions rather than exacerbating existing inequities.

1.2. ML or AI?

There is some slippage in this report between the terms machine learning (ML) and artificial intelligence (AI), though these are not synonyms. In brief, AI is the broader term, and comprises a range

[18] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, 1 edition (New York: Crown, 2016), pg. 21.

[19] Courtney Linder, "The Librarians of the Future Will Be AI Archivists," Popular Mechanics, May 13, 2020, https://www.popularmechanics.com/technology/a32436235/library-of-congress-machine-learning-newspaper-images/.

of methods and research goals, including but not limited to ML. When I discuss AI in this report, it tends to be in reference to an existing project or report that takes the wider field as its subject. This is particularly true in much of the literature focused on ethics. I cite such pieces when the points they make about AI in general pertain to ML in particular. For example, most recommendations about ethical AI are also useful guidelines for ethical ML. I will attempt to use the term "ML" exclusively in my own comments, unless I am addressing a specific document that explicitly takes a wider view of the field. In all cases where I do invoke "AI," it can be read as a rough synonym for "ML," at least in the limited context of this report.

1.3. The History of Library ML

The literature on machine learning in libraries is longer than we might imagine from the intense attention the field has received in recent years. I will not attempt to summarize it in this very brief section, but I will point to a few examples that illustrate how long library professionals have been considering the potential benefits and consequences of ML technologies. In 1976, for example, Linda C. Smith wrote in *Information Processing & Management* that the transition of information from "[b]atch tape-based retrieval systems" to "on-line retrieval systems" meant that Vannevar Bush's 1945 notion of the Memex might be within reach. Smith outlines a series of interventions ML and AI might make for information retrieval. These technologies would assist researchers, Smith imagined, with pattern recognition, feature extraction, classification, retrieval speed, information representation, problem solving and planning, and, most centrally, discovery.[20]

Two decades later, Esposito *et al* would echo these themes almost precisely in the specific context of digital libraries:

> Machine learning, together with intelligent object-centered techniques, can offer a valuable support when building intelligent digital libraries. Indeed, all the tasks related to information capture and semantic indexing can take advantage of the use of intelligent techniques and machine learning methods for layout analysis, document classification and understanding, while the integration of worldwide distributed digital libraries demands the definition of a standard query language for information retrieval. Moreover, machine learning techniques allow to infer [sic] user models from user interactions and this turns out to be useful to implement an adaptive interface.[21]

In the same period, researchers such as Chen *et al* experimented with ML to enrich geographic information retrieval in digital library systems to include metadata "that is both fuzzy and concept-based."[22] I cite only three examples over decades of work only as a gesture toward the kinds of hopes scholars have invested in ML techniques for libraries and other cultural heritage institutions.

[20] Linda C. Smith, "Artificial Intelligence in Information Retrieval Systems," *Information Processing & Management* 12, no. 3 (January 1976): 189–222, https://doi.org/10.1016/0306-4573(76)90005-4, pg. 194-215.

[21] Floriana Esposito et al., "Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service," *International Journal on Digital Libraries* 2, no. 1 (October 1, 1998): 3–19, https://doi.org/10.1007/s007990050033, pg. 18.

[22] Hsinchun Chen et al., "A Geographic Knowledge Representation System for Multimedia Geospatial Retrieval and Analysis," *International Journal on Digital Libraries* 1, no. 2 (September 1, 1997): 132–52, https://doi.org/10.1007/s007990050010, pg. 132.

Such hopes resonate across at least four decades of research in ML and libraries, centered around the potential for computer-assisted description, organization, and discovery across massive and rapidly-growing digital collections.

Despite this long history, the conversation about machine learning in cultural heritage sectors has become more urgent in the past decade because the costs of necessary hardware have fallen dramatically while ML software tools have become more readily available, meaning a much wider range of scholars and institutions can experiment with ML and potentially incorporate its methods and resulting data into their work. At the same time, widespread use of ML in technological, corporate, and governmental sectors has amplified urgent questions about the ethics of ML. Both technically and socially, then, we find ourselves in a pivotal moment, one which would benefit greatly from the expertise of librarians, curators, academic researchers, and other cultural heritage professionals.

1.4. Report Process and Scope

This report was researched and written at the request of the Library of Congress's LC Labs, following the Machine Learning + Libraries Summit held on September 20, 2019. The summit brought together librarians, computer scientists, humanities researchers, engineers, technologists, and data scientists to present state-of-the-art machine learning research; discuss best practices for organizing and conducting machine-learning research in cultural heritage sectors; and outline an agenda for the next six months, two years, and five years of machine learning research in libraries. In addition to the presentations, conversations, and post-event reporting related to the LoC summit, I undertook a thorough survey of machine learning and libraries literature, included as a bibliography in Appendix A; reviewed a range of machine learning projects, past and present; and interviewed librarians and scholars unable to attend the event. A list of those interviewees can be found in Appendix B.

I do not imagine this report to be a comprehensive account of machine learning's history in relationship to libraries, or an index of all pertinent research papers or projects. While I have made a good faith effort to survey the field, no doubt work has been missed, and my own geographic, disciplinary, racial, and linguistic perspectives have influenced the scope of this overview. I am an academic researcher whose work is deeply invested in computational analysis of library collections, and no doubt that perspective shapes the priorities outlined here, as well as the limitations of the report's view. I am a white man—a vastly over-represented category in ML research both inside and outside libraries—and while I have worked to include vital scholarship advocating for racial and cultural justice in ML, I am sure those efforts are inadequate. I urge readers of this report to seek out the work on justice I cite, read it, and then read the work cited in that scholarship as well. I hope other researchers will see this document as an invitation to further discussion, amendment, and addition.

Instead of striving for comprehensive coverage, the report aims to overview the field's major questions, cautions, and opportunities, and to advance a set of practical recommendations for cultural heritage institutions seeking to undertake ML experiments. Finally, I should emphasize that this is not a technical document, though it includes many in its citations and bibliography. Instead, I conceived and attempted to write this report primarily as an act of translation for librarians, schol-

ars, and administrators who may or may not (yet) understand the technical processes underlying ML work.

1.5. ML + Libraries Summit Event Summary

The Machine Learning + Libraries Summit that catalyzed this report was documented in careful detail by the LC Labs staff, who reported the research talks during the event, as well as gathered, transcribed, and synthesized a range of artifacts produced by attendees during group work sessions. This extensive documentation was gathered in the event report, "Machine Learning + Libraries Summit Event Summary," written by Eileen Jakeway, Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, and Abbigail Potter and published in February 2020. This event summary can be found online and may be appended to the PDF version of this report as Appendix D.[23]

1.6. Parallel Reports

This report appears in close proximity to several others, which I briefly describe below. While academic reports or articles often appear in tandem, particularly around current, pressing issues in a field, such temporal nearness can result in an odd mutual silence. That will not be the case in this report. Instead, I have learned much from my colleagues' work while preparing this document, and I hope to echo and amplify their most urgent recommendations. I strongly urge readers to consult each of the reports referenced below for a fuller sense of the current conversation around ML and libraries.

1.6.1. Responsible Operations

This report appears only a few months after Thomas Padilla's OCLC Research Position Paper, "Responsible Operations: Data Science, Machine Learning, and AI in Libraries" (henceforward RO).[24] Padilla's account of ML for libraries is nuanced, thorough, and strongly attuned to the moral and ethical implications of ML methods for cultural heritage practitioners and institutions. In particular, I admire RO's refusal to separate the moral and ethical questions surrounding ML from the practical and technical details of its implementation, a priority the report establishes from its outset:

> [T]he agenda joins what can seem like disparate areas of investigation into an interdependent whole. Advances in "description and discovery," "shared methods and data," and "machine-actionable collections" simply do not make sense without engaging "workforce development," "data science services," and "interprofessional and in-

[23] Eileen Jakeway et al., "Machine Learning + Libraries Summit Event Summary" (LC Labs Digital Strategy Directorate, February 13, 2020), https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

[24] Thomas Padilla, "Responsible Operations: Data Science, Machine Learning, and AI in Libraries," OCLC RESEARCH POSITION PAPER (Dublin, Ohio: OCLC Research, December 9, 2019), https://www.oclc.org/content/dam/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.pdf.

terdisciplinary collaboration." All the above has no foundation without "committing to responsible operations."[25]

I have interviewed Padilla since his report appeared and my work owes much to his insight and intellectual generosity. Practically, this report will refer to Padilla's throughout and seek to amplify many of his report's recommendations, while taking its overall call for responsible operations as a mandate.

1.6.2. Digital Libraries, Intelligent Data Analytics, and Augmented Description

This report also appears shortly after "Digital Libraries, Intelligent Data Analytics, and Augmented Description," written by Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack, who represent the Aida digital libraries research team (henceforward "IDA").[26] The IDA report grows out of a collaborative project between the Aida team and the Library of Congress to "develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery" and help the Library plan social and technical infrastructure to support future such projects.[27]

The IDA report focuses squarely on the Library of Congress in its recommendations, which nonetheless have broad applicability across the communities this current report addresses. IDA offers two overarching recommendations, that "the Library should focus the weight of its machine learning efforts and energies on social and technical infrastructures for the development of machine learning in cultural heritage organizations, research libraries, and digital libraries" and also "invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections." From here, they outline six more specific recommendations, urging the LoC to:

> (1) Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage. (2) Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to the larger cultural heritage and other research communities. (3) Focus efforts on developing ground truth sets and benchmarking data and making these easily available. Nested under the recommendation to support ongoing explorations and investigations, we recommend that the Library: (4) Join the Library of Congress's emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate. (5) Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas. (6) Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine

[25] Padilla, pg. 6.

[26] Prepared Elizabeth Lorang et al., "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project," January 10, 2020, 43.

[27] "Image Analysis for Archival Discovery (Aida)," 2020, http://projectaida.org/.

learning explorations.

I quote this summary of IDA's recommendations in full because they inform many of my own in this report. As with RO, whenever possible I seek to amplify many of IDA's messages, and urge readers of this report to consider IDA in full as well.

1.6.3. AI Museum Planning Toolkit

Machine learning is only one facet of the wider set of approaches designated as Artificial Intelligence, or AI (see 1.2), but the major questions and concerns around AI pertain to machine learning discussions. "AI: A Museum Planning Toolkit" was written by Oonagh Murphy (Goldsmiths, University of London) and Elena Villaespesa (School of Information, Pratt Institute) following a wide range of "conversations, workshops, and public events" (henceforward "AIT"). According to its authors, AIT:

> distils some of these conversations, flags areas for critical engagement, and serves as a practical starting point for museum professionals who are interested in working with technologies that fit within the broad field of Artificial Intelligence. The aim of this toolkit is to support non specialists to better understand the possibilities of these technologies, and empower a wide range of museum professionals to develop - strategically, ethically, and operationally robust project plans.[28]

AIT comprises a description of AI experiments underway the American Museum of Natural History and Metropolitan Museum of Art in New York City and The National Gallery in London; a worksheet meant to guide museums considering an AI project through essential questions; an AI ethics workflow; a stakeholders management worksheet; and a glossary of common AI terms.

I especially admire the practical documents developed for the AI Toolkit, such as its worksheets and workflow diagrams. The structures of worksheet and diagram help ground conversations that could spin off into abstraction in the concreteness of responsibilities, reporting, iteration, and accounting. In the current report I seek to mirror this approach, particularly in Section 6, "Questions to Guide Structuring an ML Project."

1.6.4. Principled Artificial Intelligence

In January 2020, a group of researchers from Harvard's Berkman Klein Center for Internet and Society released "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI" (henceforward PAI) a report that "analyzed the contents of thirty-six prominent AI principles documents, and in the process, discovered thematic trends that suggest the earliest emergence of sectoral norms."[29] The eight themes PAI identifies across the reports it

[28] Oonagh Murphy and Elena Villaespesa, "AI: A Museum Planning Toolkit" (Goldsmiths, University of London, January 2020), https://themuseumsainetwork.files.wordpress.com/2020/02/20190317_museums-and-ai-toolkit_rl_web.pdf, pg.1.

[29] Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," *SSRN Electronic Journal*, 2020, https://doi.org/10.2139/ssrn.3518482, pg. 4.

summarizes are privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values.

Each of PAI's themes includes a number of subtopics, many of which will be described in later sections of this report, where they most closely align or overlap the discussion of machine learning in libraries. Importantly, the authors of PAI note, "**that more recent documents tend to cover all eight of these themes**, suggesting that the conversation around principled AI is beginning to converge, at least among the communities responsible for the development of these documents." The existence of the reports I list in this section signal, perhaps, that a similar convergence is underway for discussions of machine learning in libraries, though this domain does not have the critical mass evidenced in the communities surveyed by PAI.

1.7. Acknowledgements

This report only exists thanks to the generous contributions of participants at the Machine Learning + Libraries Summit (see a list of presenters in Appendix C) and the colleagues who donated their time and mental labor for interviews (see a full list of interviews in Appendix B). I especially thank the LC Labs team, who organized the event, gathered and synthesized the artifacts produced by participants, and provided herculean support throughout the preparation of this report. In particular, Meghan Ferriter, Abigail Potter, Eileen Jakeway, and Kate Zwaard read drafts and provided advice that drastically improved the final document. I also want to thank my family, especially Evie Cordell, who helped facilitate a much-revised writing regimen in isolation during a pandemic, as we were all trying to do our jobs and/or go to school in a too-small space.

2. Machine Learning Cautions

The literature on machine learning and libraries attends far more commonly than the broader ML literature to the ethical concerns raised by ML methods. Nevertheless, much work remains to be done to help librarians and scholars consider the ramifications of ML experiments, particularly for marginalized communities. Thomas Padilla's RO report offers the most comprehensive overview of these cautions to date, emphasizing that "the challenge of doing this work responsibly requires fostering organizational capacities for critical engagement, managing bias, and mitigating potential harm" (for more on RO see 1.6.1).[30] This report likewise foregrounds these critical issues to emphasize that pragmatic questions of ML implementation must follow ethical questions about the potential effects of its implementation to the communities cultural heritage institutions represent and serve.

Padilla's report draws from and contributes to the broader work underway to interrogate algorithmic bias and injustice. As libraries design ML projects, they must first attend to scholarship, such as Safiya Umoja Noble's demonstration of how "automated" processes, such as search results based

[30] Padilla, "Responsible Operations," pg. 6.

in part on user input and behavior, can reinforce and reify white supremacy;[31] Cathy O'Neil's examples of algorithms designed to more efficiently determine creditworthiness or recidivism risk, but which in fact exacerbate racial and socioeconomic inequality;[32] and Virginia Eubanks' arguments that automated systems, such as those managing policing or resource allocation, disproportionately hurt the poor, members of the working class, and people of color.[33] Computational systems always come laden with their creators' biases and oversights. Whether purposefully or inadvertently, software designed to create business or government efficiencies often reinscribes structural inequalities.

The consequences of such systems might seem less immediately harmful when applied in cultural heritage domains, but that does not absolve libraries, universities, museums, and similar institutions from responsibility. As institutions that serve the public and scholarly communities, libraries have a unique duty to lead the conversation about ethical ML rather than following the same reckless paths we observe in the technology sector. The library ML conversation can build on the long-delayed, but necessary dialogues now underway around ethical archiving to ensure that library ML experiments intentionally address, rather than reinscribe, the oversights and violence of historical collection and descriptive practices. Libraries employ many experts in data design who are especially alert to issues of representation and social justice and who prioritize the creation of just systems. Libraries can support this expertise, designing ML projects that make unique and essential contributions to broader academic and social conversations around ML and algorithmic justice.

The point of this cautions section is not to throw our collective hands up, or to stymie ML work in libraries entirely. If libraries hope to serve all of their patrons, then these critiques must be incorporated into their ML research. Acknowledging and responding to critiques need not lead to stagnation. Instead, by taking this scholarship seriously, libraries can catalyze a unique role for themselves and their work within the broader cultural conversation about ML. D'Ignazio and Klein's book *Data Feminism*, which I discuss in more detail below, does not position itself as a call to resignation, but to action: "Data are part of the problem, to be sure. But they are also part of the solution…the power of data can be wielded back."[34] Libraries and cultural heritage organizations can be central to reclaiming data and ML methods for the communities they serve and modeling a more equitable, just vision of ML in society.

2.1. Managing Bias

Algorithms are neither neutral nor objective. They are programmed by human beings, who have both conscious and unconscious biases, and those biases are encoded into software. Usually—though certainly not always—those biases are programmed inadvertently, but they have real and sometimes devastating, consequences either way. We might be tempted to view ML—and particularly unsupervised ML—as more neutral than other algorithms because its pipeline is not laid out

[31] Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, 1 edition (New York: NYU Press, 2018).

[32] O'Neil, *Weapons of Math Destruction*.

[33] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York, NY: St. Martin's Press, 2018).

[34] Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (Cambridge, Massachusetts: The MIT Press, 2020), pg. 17.

in advance; we might be tempted to believe an unsupervised ML process merely reveals patterns inherent in data. However, datasets are no more neutral or objective than algorithms. They are constructed by human actors to meet particular ends, whether political, social, or academic. Some facets of experience are recorded while others are not, and those recording data are not always fully cognizant of how their choices are shaped by their own biases.

What libraries choose to digitize often reflects the biases and priorities of scholars and users from dominant groups over others, and as a result our current digital collections should not be considered comprehensive or representative. As the IDA report emphasizes,

> Previous and ongoing collecting and description practices, for example, were and are colonialist, racist, hetero- and gendernormative, and supremacist in other structural and systemic ways. These understandings are the foundation on which training and validation data will be created and assembled; they will become reinscribed as statements of truth, even as we elsewhere champion the potential of computational approaches to uncover hidden histories, identities, and perspectives in collections. To engage machine learning in cultural heritage must mean confronting these histories, committing to the hard work of acknowledgment and rectification, and not simply reproducing them and giving them a whole new scale of power.[35]

Here IDA reminds us that, as with scale, problems of representation in library collections are older than digitization or ML, though these new technologies threaten to reify past biases if not explicitly conceived and implemented toward justice.

In her book *Algorithms of Oppression*, Safiya Umoja Noble demonstrates how the data underlying internet behemoths such as Google skew search results, such that searches for images of Black teenagers reinforce the racial stereotypes of white users, while searches for white teenagers result in largely anodyne results. As Noble shows, such search results are not rare mistakes, but common testaments to systemic problems in the datasets from which Google's search engine learns.[36] Whether such biases were intended is irrelevant; the resulting harms are real. In her book *Weapons of Math Destruction*, Cathy O'Neil shows similar biases at work in existing systems for determining creditworthiness for consumers or the likelihood of recidivism for convicted criminals. While such systems do not explicitly factor race into their calculations, they strongly weight data points that stem from racist practices. Those biased algorithms amplify the racist data on which they are based and magnify its effects going forward, meaning in practice that Black applicants are less likely to be extended credit, while Black people convicted of crimes are more likely to be denied bail or given more severe sentences than white people convicted of the same crimes.[37] These examples show that creating or using data to train machine learning algorithms are not neutral steps and can have unintended consequences for how libraries provide service to their users.

The ML processes described by scholars such as Noble and O'Neil have immediate, practical consequences to people's lives and livelihoods. By contrast, the materials in libraries and other cultural heritage institutions might seem less immediately pressing, or less actively harmful. However, the data available in library collections will determine the range of possibility for any ML projects un-

[35] Lorang et al., "Digital Libraries, Intelligent Data Analytics, and Augmented Description," pg. 28.
[36] Noble, *Algorithms of Oppression*.
[37] O'Neil, *Weapons of Math Destruction*.

dertaken. The biases, limitations, and oversights of those datasets will produce flawed research that does not represent the communities libraries seek to serve. An ML project that uses library subject terms to categorize data, for example, may inadvertently echo and amplify outdated or even offensive subject terms. In Section 4.1.3 I will describe how such terms can be found in the tags underlying the ImageNet library, widely considered a "gold standard" training dataset for image-based ML projects. Catalogers are working hard to update such terms in library metadata systems, but this process will be slow and ongoing, and ML researchers must be cautious about assuming legacy systems are appropriate for ML training.[38]

As Catherine D'Ignazio and Lauren Klein argue in *Data Feminism*,

> The problems of gender and racial bias in our information systems are complex, but some of their key causes are plain as day: the data that shape them, and the models designed to put those data to use, are created by small groups of people and then scaled up to users around the globe. But those small groups are not at all representative of the globe as a whole, nor even of a single city in the United States. When data teams are primarily composed of people from dominant groups, those perspectives come to exert outsized influence on the decisions being made—to the exclusion of other identities and perspectives.[39]

While perhaps not as skewed as the demographics of Silicon Valley tech companies, nonetheless the demographics of libraries—and particularly library technical staff—and academic researchers do not well reflect the US population. These groups are predominantly white, for instance, and that fact shapes what materials exist in archives, which of those materials are chosen for digitization, and the computational systems—including ML—employed to analyze those digital resources.

Bias can be introduced into ML algorithms at many critical points; one of these has to do with the collection and organization practices underlying the data itself. Users of large data sets, especially those developed in long-term projects, should consider the limitations and biases of the data before developing ML projects. For example, the Chronicling America newspaper collection, developed as part of a five-decade long initiative to preserve historic U.S. newspapers, provides one of the largest corpuses of freely-available digitized newspaper data. While this makes the project extremely attractive to developers of humanities-related machine learning projects, the data skews to newspapers serving the majority. For instance, researcher Ben Fagan has shown how the collection privileges newspapers that served the dominant white, middle-class readers of the nineteenth century, to the near exclusion of Black and other minority-run papers.[40] This reality was the not the result of any concerted effort to exclude minority papers, but resulted from state-level digitization efforts prioritizing geographic representation, for example by choosing to digitize one paper from each of their state's most populous cities.[41] By prioritizing geographic spread, most states inadvertently deemphasized racial or cultural representation, and consequently any ML projects based on

[38] Anna M. Ferris, "The Ethics and Integrity of Cataloging," *Journal of Library Administration* 47, nos. 3-4 (July 1, 2008): 173–90, https://doi.org/10.1080/01930820802186514.

[39] D'Ignazio and Klein, *Data Feminism*, pg. 28.

[40] Benjamin Fagan, "Chronicling White America," *American Periodicals: A Journal of History & Criticism* 26, no. 1 (2016): 10–13, https://muse.jhu.edu/journals/american_periodicals/v026/26.1.fagan.html.

[41] Ryan Cordell, ""Q I-Jtb the Raven": Taking Dirty OCR Seriously," *Book History* 20 (2017): 188–225.

Chronicling America will reflect those same oversights and exclusions.[42] While state partners have sought to be more representative in the decades since the program began, an understanding of the underlying collection is critical to any ML projects undertaken using this archive.

As libraries develop ML plans and projects, the current state of datasets cannot be dismissed as a regrettable but unavoidable reality. Instead, libraries must prioritize the construction of equitable datasets. As Padilla notes in RO, eliminating bias is not a feasible goal for library ML projects. Instead, the RO report advocates for managing bias:

> Managing bias rather than working to eliminate bias is a distinction born of the sense that elimination is not possible because elimination would be a kind of bias itself—essentially a well-meaning, if ultimately futile, ouroboros. A bias management paradigm acknowledges this reality and works to integrate engagement with bias in the context of multiple aspects of a library organization. Bias management activities have precedent and are manifest in collection development, collection description, instruction, research support, and more.[43]

RO's recommendations for managing bias require self-conscious attention and communication at every stage of an ML project, from symposia and intellectual exchanges about collection biases during projects' development stages to audits of projects' results at multiple checkpoints.

As D'Ignazio and Klein argue, the very goal of mitigating bias might not go far enough, as terms such as "bias," "ethics," "fairness," or "accountability" all "locate the source of the bias in individual people and specific design decisions" rather than in systems of power. They advocate "A broader focus on *data justice*, rather than *data ethics* alone" to "ensure that past inequities are not distilled into black-boxed algorithms." Following the very useful table provided in *Data Feminism* for moving from "data ethics" toward "data justice," library ML projects should not seek to simply mitigate *bias*, though that may be a useful starting point, but instead to actively challenge *oppression*. In considering the collections that underly ML projects, libraries should not seek to simply encode an idea of *fairness*, but instead to work actively toward *equity*.[44] To create ML projects that reflect data justice, in other words, libraries cannot pretend to be neutral or objective in relationship to race, class, gender, sexuality, or culture, but instead must consciously strive to forefront marginalized voices.

Data ethics and justice are growing areas of interest in computer science, exemplified by the work of scholars such as Rediet Abebe and Emily Denton, whose work libraries might follow to develop projects that do not acquiesce to the fallacy that biased datasets must result in biased ML research. In a paper from earlier this year, Abebe *et al* argue that too much technical work claiming to mitigate bias "treats problematic features of the status quo as fixed, and fails to address deeper patterns of injustice and inequality"— thus amplifying the very biases the work seeks to address. Like D'Ignazio

[42] I include in this group the Viral Texts project (https://viraltexts.org/), on which I am co-PI. While we have worked to expand our corpora to address the limitations of representation in Chronicling America, we have not yet done enough to work actively toward equity. The analyses we have produced have taught us more about nineteenth-century newspaper editors and readers, but predominantly about white, middle-class editors and readers rather than the many other populations of editors and readers in the period.

[43] Padilla, "Responsible Operations," pg. 9.

[44] D'Ignazio and Klein, *Data Feminism*, pg. 60-61.

and Klein, this paper argues that computational research can help address social problems, and articulates four key modalities for doing so:

> Computing research can serve as a *diagnostic*, helping us to understand and measure social problems with precision and clarity. As a *formalizer*, computing shapes how social problems are explicitly defined—changing how those problems, and possible responses to them, are understood. Computing serves as *rebuttal* when it illuminates the boundaries of what is possible through technical means. And computing acts as *synecdoche* when it makes long-standing social problems newly salient in the public eye.[45]

These four modalities hold enormous potential for libraries undertaking ML research, and could provide motivating intellectual structure where we might be otherwise tempted to stagnation. Rather that trying to consider every ML project in terms of fairness, a library ML project might instead serve as a *diagnostic* to the problems inherent in existing digital collections, a *rebuttal* to ML work falsely claiming objectivity, or a *synecdoche* that helps patrons better understand the historical stakes of library collections and archives. These three aims meet the pedagogical goals for library ML I outline in 5.5.1, while also aligning with some of the most vital ML research underway in computer science.

2.2. Participant Consent

Though these themes are less prominent in the literature than research focused discussion, certainly some in the library and cultural heritage community see potential in ML methods for better understanding patrons: both who they are and what they are seeking. The AIT, for instance, presents a case study in which sentiment analysis was performed on visitor surveys and comments in order to better understand those documents *en masse*.[46] In the library context, some imagine experiments that, for example, analyze patrons' borrowing habits in aggregate to better plan future collection development. A more extreme proposal might be to employ facial recognition to identify patrons for the purpose of using library resources, or checking out materials.

Developments like these raise especially urgent ethical questions, particularly around consent. Libraries should take warning from the mistakes that have proliferated in the technological and corporate spheres in the past years and proceed with extreme caution if considering the implementation of patron-focused ML. I take note, for example, that among the many reforms responding to `#BlackLivesMatter` protests in spring 2020, IBM announced the company "will no longer offer general purpose facial recognition or analysis software" and will also no longer develop or research the technology" after recognizing the extent to which such technologies perpetuate and amplify racial bias.[47] Libraries have traditionally supported privacy and anti-surveillance measures and

[45] Rediet Abebe et al., "Roles for Computing in Social Change," January 28, 2020, https://doi.org/10.1145/3351095.3372871.

[46] Murphy and Villaespesa, "AI," pg. 4.

[47] Jay Peters, "IBM Will No Longer Offer, Develop, or Research Facial Recognition Technology," The Verge, June 8, 2020, https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-recognition-analysis-software.

again with ML-enabled surveillance systems, libraries can act slowly and deliberately, ultimately refusing to implement technologies that have the potential to harm or marginalize patrons.

One general principle of ethical ML/AI is that people affected by the results of an algorithmic process must be made aware of that process: including its source data, the ways that data is weighed and evaluated, and the steps through which a transformation or decision is enacted. But simple awareness is not sufficient. If people's data is to be used by an ML process, they must first consent to such use. Given their public charge and mission, libraries must take these charges seriously and respond proactively, rather than reactively, to imperatives of notification and consent. Libraries' traditional commitment to protect the privacy of patrons certainly carries forward into any ML systems they consider employing. At the very least, libraries must commit to keeping any patron data emerging from an ML project private and restrict any commercialization of that data.

Going further, any ML system that directly affects patrons must be fully documented and advertised across the media that different communities of patrons will encounter. If a community of patrons is unlikely to see a notice on the library's website, for example, that cannot be the only medium through which the information is shared. In addition, such systems should be deployed as "opt in" rather than "opt out," meaning that patrons should be given agency to decide whether to participate, not forced to decide to withdraw from a system already implemented. Finally, any patron-focused ML systems must be subject to regular audit and re-evaluation by both library staff and the affected community; see 5.1.6 for recommendations around algorithmic auditing.

2.3. Environmental Impact

The literature on ML in libraries—frankly, like the broader ML literature—has not sufficiently grappled with the environmental effects of the technology, which can be substantially more onerous than typical computing (which is itself not immune from environmental critique). In a 2018 study, updated in 2019, a group from OpenAI found an exponential increase over time in the amount of compute used to train AI models. Looking at data from 1959 to 2019, they found "two distinct eras of training AI systems in terms of compute-usage: (a) a first era, from 1959 to 2012, which is defined by results that roughly track Moore's law, and (b) the modern era, from 2012 to now, of results using computational power that substantially outpaces macro trends."[48] While many researchers hope ML and AI will help researchers better identify ways to model and mitigate climate change, the technologies themselves have environmental consequences that must be considered.

In a 2019 paper, Strubell *et al* found that training a deep neural network model "incurs a substantial cost to the environment due to the energy required to power this hardware for weeks or months at a time," while most of the power current used to train such models does not come from renewable sources. Most dramatically, this study finds that training "the most computationally-hungry models" can emit as much CO2 as five cars would over their entire lifetimes. This study found that "model training and development likely make up a substantial portion of the greenhouse gas emissions attributed to many NLP researchers," which does point to the need for models that can be useful across experiments, projects, or institutions.[49]

[48] Dario Amodei et al., "AI and Compute," OpenAI, November 7, 2019, https://openai.com/blog/ai-and-compute/.
[49] Emma Strubell, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in

Such resource-intensive model training is called "Red AI," and can be contrasted with "Green AI," or "AI research that yields novel results without increasing computational cost, and ideally reducing it."[50] Schwartz *et al* advocate for more widespread Green AI practices in which efficiency becomes a key criterion for evaluating AI research and researchers are ethically bound to report the computational and environmental resources their research requires. One key insight of Schwartz *et al* study is that researchers' lack of communication about the precise parameters and iterations of their experiments, prevent comparisons that could reduce the field's environmental impact substantially. They suggest that studies into optimal sizes for training data could ameliorate the need for very massive studies that consume significant amounts of energy, particularly if much smaller datasets yield similar results.

Similar kinds of communication, data sharing, and establishment of best practices could help the ML and libraries community reduce the environmental impact of ML projects. Given the relative newness of ML work in libraries, the field could establish early guidelines for environmental accountability that reduce the need for interventions later. Indeed, the library community could become an exemplar for ML research in other sectors by building measures for environmental accountability into its core practices and holding its community members responsible.

3. Promising Machine Learning Applications

Though attention to the field has intensified in recent years, researchers have already undertaken a wide range of ML experiments in libraries (or using library collections), such that no report could fully summarize them. In this section I seek to outline the most common applications of ML found in the current literature and define the broad areas of greatest promise for work in the immediate future. These areas should not be considered comprehensive of all ML work that might be undertaken in library collections. The future of ML work in libraries will require asking:

1. What under-explored datasets would benefit most from ML approaches and
2. What patterns within those datasets might be most meaningful to surface?

The answers to both of those questions will require creativity in both subject expertise and the application of computational methods.

3.1. Crowdsourcing

Even outside of ML contexts, library crowdsourcing projects are often intended to enhance discoverability, which will be the primary topic of 3.2. I call out crowdsourcing separately, however, as the ways libraries involve communities in annotating collections vary widely, and typically precede the ML experiments described throughout the rest of this section. Sometimes crowdsourced annotations are solicited explicitly as ML training data, while other times ML researchers find data annotated for other purposes useful with an ML context. As noted in 4.1.2, however, one of the

NLP," June 5, 2019, http://arxiv.org/abs/1906.02243, pg. 1-2.

[50] Roy Schwartz et al., "Green AI," August 13, 2019, http://arxiv.org/abs/1907.10597, pg. 1, 5.

largest challenges facing library ML work is the labor required to create meaningful training data, and crowdsourcing efforts hold much potential for addressing that need.

Meghan Ferriter from the Library of Congress points to prominent library-based crowdsourcing projects created for text-generation and classification, such as OCR correction in Trove's historical newspaper collection, which both seeks to improve the results of an ML process (OCR) for current discovery, while contributing to the development of better models for future OCR work.[51] Trove's "text correction hall of fame" celebrates the work of its community members on behalf of Australia's historical record. The Library of Congress's Beyond Words project charged volunteers with improving metadata about images in the Chronicling America newspaper corpus by drawing boxes around identified images, transcribing captions and artist names, tagging images by category, and verifying the work of other volunteers.[52] This work both assists with existing discovery interfaces, but has also contributed to training data for ML projects such as IDA and Newspaper Navigator, both of which this report cites as models for library ML. Such efforts can pay dividends across disciplines as well. For example, the Smithsonian Institution's crowdsourcing efforts to transcribe labels on digitized bee specimens helped researchers train deep learning models to "classify specimens to subgenus…and species."[53]

The appeal of crowdsourcing for ML projects is clear: if a library can mobilize a community of interested participants, it can build sufficient training data quickly, even in domains for which little or no training data exists. This can save both time and money while generating patron interest in the project's outcomes. However, libraries cannot expect the movement from crowdsourcing to ML to be seamless. As Ferriter emphasized, a "required condition…for engaging with crowdsourced data as training data for ML" is having a "thorough understanding of the crowdsourcing project's goals, participants, management and significantly its workflow and data models." Focusing on the crowdsourcing for Beyond Words, Ferriter notes that the original goals of the project were not to generate ML data, and thus

> the nuances of interpreting and legibility of images (grainy photos can look like illustrations), barriers created by the project workflow, and consensus seeking "goals" of the system—coupled with no community management (intentional)—have a great impact on the resulting quality and/or accuracy of the data.[54]

In addition, libraries should be cautious of an "if we build it, they will come" model of crowdsourcing. Successful crowdsourcing projects typically comprise a relatively small group of dedicated contributors and a penumbra of users making less sustained contributions. That core of dedicated users may be drawn from a scholarly organization with an intellectual investment in the project's outcomes, a community whose materials comprise the project's data, or a group of hobbyists, but their work must be motivated and recognized, and they must be informed about any uses, including ML, to which their data is being applied.

---

[51] Trove, "Text Correction Guidelines," accessed May 21, 2020, https://help.nla.gov.au/trove/digitised-newspapers/text-correction-guidelines.

[52] Library of Congress, "Beyond Words," 2020, http://beyondwords.labs.loc.gov/#/.

[53] Chandra Earl et al., "Discovering Patterns of Biodiversity in Insects Using Deep Machine Learning," *Biodiversity Information Science and Standards* 3 (July 2, 2019): e37525, https://doi.org/10.3897/biss.3.37525.

[54] Meghan Ferriter, "Report Draft Comments," May 15, 2020.

Patrons who contribute their labor to research must know they are doing so and willingly consent, and the burden is on libraries to construct fair systems for such participation. Libraries should be critical of efforts that rely, for example, on uncompensated student labor through assignments. As Spencer Keralis reminds us, crowdsourcing "relies on a social contract of volunteerism and informed consent that is simply not possible in the classroom. Student labor in the classroom is never not coerced…The power dynamic of the classroom is such that student choice in this situation cannot be unequivocal, and that faculty objectivity will always be suspect."[55] Crowdsourcing can be considered another patron-focused application of ML, and as such requires the same proactive communication and consent-seeking as other ML applications.

3.2. Discoverability In and Across Collections

By far the most widely-cited application for machine learning in libraries is to aid discovery in large-scale collections. Experiments have demonstrated the usefulness of ML for extracting machine-actionable text from digitized documents; identifying common topics or ideas across documents; extracting metadata from digitized objects; tagging the content of digital collections; grouping similar materials, whether textual, visual, or auditory; or identifying geographic features in maps and related materials. The most widespread ML technology in libraries, Optical Character Recognition (OCR), already aids search and discovery across many library collections, though libraries have been slower to integrate the results of other ML techniques into their interfaces and catalog systems.

3.2.1. Clustering and Classification

ML methods are often cited as valuable because of their ability to identify connections between materials that might not be apparent to humans browsing collections, due to the limitations of human attention and memory, or even within curated metadata such as catalog records. As large collections of media coalesce in digital forms, including collections whose originals are physically dispersed, libraries look to ML to help group similar materials and enable paths for serendipitous discovery. Through a supervised classification experiment, a corpus of texts could be sorted by genre, theme, or structure, while unsupervised ML experiments could sort the same texts by shared topics or linguistic structures.

Clustering and classification experiments have applications both for domain research and for library systems. For example, a team of researchers trained a classifier on the knowledge classes in the eighteenth-century *Encyclopédie* in order to evaluate the internal consistency of the original editors' ontologies.[56] In this case, the purpose of the research was to better understand the source material: to learn about categories of knowledge in the eighteenth century. A more recent example from the "Living With Machines" project team, housed in the British Library's Alan Turing Institute, used contextualized word embeddings to detect moments of "atypical animacy" in nineteenth-century

[55] Spencer D. C. Keralis, "Milking the Deficit Internship," in *Disrupting the Digital Humanities*, ed. Dorothy Kim and Jesse Stommel, 2016, http://www.disruptingdh.com/milking-the-deficit-internship/.

[56] Russell Horton et al., "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie," *Digital Humanities Quarterly* 003, no. 2 (June 18, 2009).

writing—in other words, to find passages in a corpus when machines are represented as living beings.[57] Other classification experiments have focused on contemporary knowledge organization. A team of researchers from the National Library of Norway describe an experiment that uses ML methods to automatically group articles and assign Dewey Decimal numbers to them, to aid in cataloging.[58]

ML methods are not restricted to textual materials, and in fact some of the most promising current ML experiments draw on computer vision methods to cluster visual materials, such as paintings and photographs. The "National Neighbors" project, led by researchers at Carnegie Mellon University and the University of Pittsburgh, used a convolutional neural network to relate materials from National Gallery of Art across 2,048 dimensions. The resulting visualizations allow researchers and other interested users to explore these collections by visual similarity, and such data could be incorporated into an interface so that users looking at a particular artwork would be prompted to other similar pieces. Similarly, the "Neural Neighbors" project from Yale's DH Lab used an Inception Convolutional Neural Network to estimate similarity among more than 27,000 historical photographs in their Meserve-Kunhardt Collection. When users click a particular photograph, the interface prompts them to explore the most similar photographs from the collection, each listed with a percentage indicating its similarity to the original photograph. This simple act of reporting some of the probabilities generated by the ML algorithm speaks to the issue of ML explainability, which I will discuss in 5.2.2.[59]

PixPlot, from the same research team, plots the pictures in a 3D visualization for browsing.[60] Such clouds of images recall, perhaps, the word clouds stereotypically associated with early textual analysis work, but could provide an effective alternative interface for serendipitous discovery in visual collections. For example, Liz Gallagher describes how researchers trained a neural network on the Wellcombe Collection Archive to create a "a rich 1x1x4096 feature vector" about 120,576 images from their collections. Using these vectors, they visualize the entire collection in three-dimensional space, where visually similar images appear nearby each other; measure the visual distance between images; and construct "paths" between any two images in the collection. These are essentially chains of images that constitute the shortest algorithmic distance between the two input images.[61] Such paths are interesting in their own right, but also suggest new paths of research inquiry (how does the work of artist A and B relate?) or serendipitous exploration of a collection.

As the "National Neighbors" researchers write, however, such experiments also prompt meta-questions about what similarity itself means for teaching or research, to "understand how the visual similarity derived by Inception-v3 [their chosen neural network] tracks with the formal qual-

[57] Mariona Coll Ardanuy et al., "Living Machines: A Study of Atypical Animacy," May 22, 2020, http://arxiv.org/abs/2005.11140.

[58] Svein Arne Brygfjeld, Freddy Wetjen, and André Walsøe, "Machine Learning for Production of Dewey Decimal," July 19, 2018, 9.

[59] Yale Digital Humanities Lab, "Neural Neighbors," accessed September 17, 2019, https://dhlab.yale.edu/neural-neighbors/.

[60] Yale Digital Humanities Lab, "PixPlot: Visualizing Image Fields," accessed September 17, 2019, https://s3-us-west-2.amazonaws.com/lab-apps/pix-plot/index.html?utm_source=dancohen&utm_medium=email.

[61] Liz Gallagher, "Finding Image Pathways," Medium, accessed September 17, 2019, https://stacks.wellcomecollection.org/finding-image-pathways-12d31ae347f9.

ities we'd usually associate with in art history, and how it operates differently."[62] In particular, the researchers find that their visualizations often illustrate not (only) the intrinsic visual qualities of the artworks, but the contours of a given collection versus another. In this way, ML can also help an institution understand the inclusions and oversights of its collections at scale, as well as the behavior of the ML model chosen to analyze it.

Whether the data is textual (3.2.3, 3.2.4), tabular (3.2.6), visual (3.2.7), or auditory (3.2.8), ML clustering and classification methods promise to supplement human cataloging and description to make collections more tractable through both search and browsing. By linking related materials, ML can help researchers and students explore domains more thoroughly and efficiently, aiding serendipity in digital libraries. To ensure reliability in such results, systems for iterative human evaluation of ML outputs must be established, as I discuss in 5.1.4. Moreover, the results of ML clustering and classification will always be situated and contextual, speaking only to the interrelationships of a given collection or dataset, rather than a totalizing set of relationships across a scholarly or historical domain. As Kate Zwaard noted in our interview, this work of proper contextualization is one of the single biggest challenges facing ML work, but it is one librarians in particular are well poised to undertake.[63] The expertise of librarians—particularly its archivists and catalogers—will prove necessary for contextualizing the results of ML projects within institutional histories of collecting, description, and digitization.

3.2.2. Pre-Processing

The IDA report (see 1.6.2) discusses a range of experiments undertaken by its team using newspaper data from the Chronicling America and Europeana newspaper collections. IDA separates these experiments into "first-round" and "second-round explorations." The first-round explorations they describe might be classed as infrastructural, or "pre-processing" as described in the report. IDA focuses on tasks such as segmenting documents into meaningful sectors, classifying graphic elements, extracting text, or assessing the quality of archival images. Such activities receive less attention or esteem than they deserve from humanities and DH scholars, as they perhaps seem separate from interpretation, but the more touted methods such as topic modeling, vector space analysis, image clustering, or genre classification cannot happen without these more fundamental processes. For example, effective zoning of document images leads to more effective transcription (see 3.2.3 and 3.2.4), which leads to more reliable data for historical or textual analysis.

Experimenting with a range of ML tools and algorithms, the IDA team found a number they describe as promising for tasks such as zoning or segmentation, and they also demonstrate the potential to compare pages' visual similarity (a second-round exploration). For full details about their nine explorations, see the full IDA report. Here I point to this work in order to highlight the potential for ML methods to help with necessary but labor-intensive processing for cultural heritage materials. In this vein, the IDA team "wondered about *the interplay of human expertise and processes and machine knowledge and processes*. What human-computer processes," they ask,

[62] Matthew D. Lincoln et al., "National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections," November 2019, 20, https://nga-neighbors.library.cmu.edu.

[63] Cordell, "Machine Learning and Libraries Interview with Kate Zwaard."

> might be viably and validly adopted and operationalized as, say, part of a daily routine? What human-computer approaches are viable and valid in terms of effectiveness and efficiency in order to address issues of scalability? What value might there be in cross-learning, loop-learning, and cross-processing, where machines learn from humans, humans respond to and adapt understanding based on machine learning, and this looped learning informs processes and decision-making?[64]

As libraries turn to ML to automate pre-processing steps, they must remember that pre-processing will have substantial downstream effects. Pre-*processing* does not suggest pre-*interpretation* and the structure of such work should be as carefully considered as the other tasks outlined in this section.

3.2.3. Optical Character Recognition

The most widely deployed ML technology in libraries is Optical Character Recognition (OCR), which comprises a set of algorithms designed to mimic the functions of the human eye and brain and discern which marks within an image represent letterforms or other markers of written language, transcribing these into machine-actionable text data. Typically OCR is used in situations where manual transcription would be too costly or time consuming, such as when a large corpus has been scanned. Both large- and small-scale digital archives rely on OCR, including collections such as *HathiTrust*, *EEBO*, *ECCO*, or Chronicling America. Relative to manual transcription, OCR is a quick and affordable means for creating computable text data from a large collection of images.

OCR is so widely used in digital library collections that it is easy to forget it is a ML method at all. For the most part, OCR is a supervised ML method, in which an engine is trained on a set of known documents and then used to transcribe a larger set of unknown ones. For many collections, OCR-derived data can be deeply errorful, as I will outline below, but overall I would posit the technology as a primary example of how libraries have and might continue to integrate useful probabilistic data into discovery interfaces. In short, despite its shortcomings, OCR is a ML method that has quickly become central to discovery and has, quietly but dramatically, transformed entire sectors of research.[65] Certainly scholars debate the impacts of this transformation, citing the anecdotal confirmations possible in super-abundant text archives[66] or the ways search amplifies the importance of digitized sources over undigitized ones[67], but OCR has significantly increased access to large-scale collections, including in media that were difficult to access at scale prior to its development. The use of historical newspapers, for instance, has exploded in scholarship of the past two decades, largely due to the kinds of access enabled by searchable text data, which can make vast collections tractable.

As David Smith and I outlined in our 2019 report, "A Research Agenda for Historical and Multilin-

[64] Lorang et al., "Digital Libraries, Intelligent Data Analytics, and Augmented Description," pg. 30.

[65] Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations* 127, no. 1 (2014): 64–72, http://rep.ucpress.edu/content/127/1/64.abstract.

[66] Maurice S. Lee, "Falsifiability, Confirmation Bias, and Textual Promiscuity," *J19: The Journal of Nineteenth-Century Americanists* 2, no. 1 (April 3, 2014): 162–71, https://doi.org/10.1353/jnc.2014.0014.

[67] Ian Milligan, "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 19972010," *The Canadian Historical Review* 94, no. 4 (2013): 540–69, https://muse-jhu-edu.ezproxy.neu.edu/journals/canadian_historical_review/v094/94.4.milligan01.html.

gual Optical Character Recognition" (funded by the Andrew W. Mellon Foundation and conducted in consultation with the NEH's Office of Digital Humanities and the Library of Congress),[68] there is significant opportunity for research that will create more reliable and representative OCR data over the next five to ten years. OCR was largely developed to process typewritten, English-language, mid-twentieth-century business documents. With that kind of input, OCR is remarkably reliable, transcribing with greater than 90% accuracy. When those OCR engines are applied to historical documents, however, with distinct typography, complex layouts, torn pages, smeared ink, and any number of features those OCR engines were not trained to discern, then the reliability of OCR transcription declines precipitously. A famous example of this is the German fraktur typeface, or "blackletter" in English, which was used in most German books and newspapers through the early 20th century and which OCR engines have not historically been well trained to recognize or transcribe.

For multilingual documents and languages outside English, particularly those do not use Latin script, the problem becomes even more acute and error rates significantly higher. Our report surveys a range of exciting work addressing these limitations. The *Primeros Libros de las Américas* project at the University of Puebla and Texas A&M, for example, worked to automatically transcribe books in sixteenth-century variants of Latin, Spanish, Nahuatl, and other languages—often with mixtures of these languages in, e.g., grammatical works and with significant orthographic variation within languages.[69] The *Open Islamicate Texts Initiative* (OpenITI) project focuses on building language models across Arabic, Persian, Ottoman Turkish, Urdu, and other Arabic-script languages of the Islamic world.[70] These and similar projects illustrate the enormous potential remaining for OCR research that will expand the reach and impact of digital library collections for education and research.

I will not reiterate all of our previous OCR report recommendations here, but only emphasize those that resonate with others in this report, such as the pressing need to develop better training data sets for documents from particular periods, languages, or document types. Like other ML approaches, the quality of OCR depends entirely on the training data provided, and thus the most promising developments in this area will rely on the development of more robust and diverse collections of training data. In particular, our report called for focused development in particular areas, as "knowledge domains exist for which a concerted, focused effort of funding and research could substantially advance the current state of optical character recognition for an entire subfield or research area."[71] In addition, however, we acknowledged that libraries, universities, and other cultural heritage organizations also host a wide range of transcription projects that could provide meaningful training data for improved OCR systems if systems were developed to gather and deploy them.[72]

One reason OCR is so important for ML and library research is that the availability of text data is

[68] David A Smith and Ryan Cordell, "A Research Agenda for Historical and Multilingual Optical Character Recognition," 2018, https://repository.library.northeastern.edu/files/neu:f1881m035.

[69] Hannah Alpert-Abrams, "Machine Reading the Primeros Libros" 10, no. 4 (2016), http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html.

[70] Matthew Thomas Miller, Maxim G. Romanov, and Sarah Bowen Savant, "Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans," *International Journal of Middle East Studies* 50, no. 1 (February 2018): 103–9, https://doi.org/10.1017/S0020743817000964.

[71] Smith and Cordell, "A Research Agenda for Historical and Multilingual Optical Character Recognition," pg. 25.

[72] Smith and Cordell, pg. 20-21.

often the first requirement for employing other ML techniques outlined in this report. For a researcher to create topic or vector space models of a collection, for instance, they first need machine-readable text data of that collection, which may well be created using OCR. As Smith and I outline in our OCR report, the statistical impact of errorful OCR on subsequent analytical tasks is not well enough understood, though researchers are beginning to assess it. Hill and Hengchen constructed a series of experiments to evaluate the effect of differing levels of OCR quality on a range of computational text analysis techniques, and found that OCR has less impact than anticipated on topic models, while having a larger, though not unmanageable, effect on collocations. As libraries delve further into complex ML techniques, their OCR data will become ever more important.

3.2.4. Handwriting Recognition

While there remains work to be done for OCR, even more remains for handwriting recognition, which could make libraries' substantial manuscript collections computationally tractable for full-text search, as well as the follow-on data analysis techniques described in this report. While the word "manuscript" often calls to mind medieval European texts—certainly a subset of the materials handwriting recognition would benefit—the category is much broader, including materials from around the world, and from regions or group of people for whom handwriting remained the dominant technology of textual transmission for practical or cultural reasons into the modern period. In addition, libraries hold vast collections of diaries, letters, papers, and other handwritten genres—handwritten genres which surrounded and shaped print culture throughout its history.[73] In many cases, these manuscript collections likely eclipse printed collections in both scope and representation, but remain largely inaccessible to computational analysis.

Where the regularization of print makes OCR feasible, however, the many individual nuances of handwriting have made it far more difficult to devise systems for computational detection and transcription at scale. As with other ML projects, substantial training data is required to train a model on a particular hand, but the sheer number of hands in any given manuscript collection make it impossible for most libraries or cultural heritage organizations to train models for each. Some of the most promising ML work related to handwriting recognition, then, is organized around compiling training data. The Transkribus project from the University of Innsbruck is, at base, a platform for collecting training data. Scholars using the software can transcribe primary texts (whether handwritten or printed) and contribute them to the platform, linking their transcription to the source image they used. Once a scholar has transcribed at least 100 images, the research team trains a Handwritten Text Recognition (HTR) engine on the data, which the scholar can then use to automatically transcribe more images from the same domain.

In addition to its outreach to scholars, the Transkribus project describes itself as a resource for libraries seeking to mobilize a local community of users (librarians, scholars, students) to rapidly develop an HTR model for a local collection and thereby enrich its data, as well as for computer scientists who need an environment to test tools in development.[74] As such we might look to

[73] Rachael Scarborough King, *Writing to the World: Letters and the Origins of Modern Print Genres* (Baltimore: Johns Hopkins University Press, 2018).

[74] TRANSKRIBUS Team, *Transkribus* (Innsbruck, Austria: University of Innsbruck, 2020), https://transkribus.eu/Transkribus/#archive-content.

this project as a model for motivated crowdsourced engagement (see 3.1) in order to build domain data. In 4.1.2 I describe the need among scholars and libraries for more domain-specific training data, if we hope to see a wider investment in ML projects. Transkribus provides a compelling model for community-driven efforts to create, gather, and share training data, as well as the potential for crowdsourcing for rapid dataset development. While Transkribus focuses on handwritten text recognition—alongside some OCR—its platform and community-engagement models could be adapted across a wide range of domains and ML projects.

3.2.5. Metadata Recognition and Extraction

Methods such as named-entity, location, and date recognition can be used to extract valuable metadata from digital holdings and enrich collections for browsing, searching, and computational analysis. Such methods attempt to automatically recognize meaningful words such as proper names, geographic names, and temporal markers from text data (perhaps text data extracted from images through an earlier workflow). While straightforward entities can be identified by cross-referencing dictionaries of meaningful terms, ML methods can be employed to ascertain less-well defined entities using linguistic context, document structure, or probabilistic measures of similarity. In our interview, Digital Humanities Librarian Amanda Rust cited identifying people and places within unstructured text data as one of the most exciting possibilities for ML work, particularly for historical materials and research.[75] The goal of these ML methods is to identify a particular, meaningful type of metadata within a larger, possibly unstructured dataset.

Consider work on maps and other geographic materials. More than a decade ago, Chen *et al* described an approach to inter-related resources such as "text-based geographic literature; technical reports and surveys; maps, aerial photos, satellite images, and digital elevation models; and a large variety of scientific datasets" that seeks to identify geographic features across these sources and create a "geographic knowledge representation system" that will support "concept-based access" to those collections.[76] The goal of this research was to identify meaningful features and link related material across media, so that collection users could trace paths not apparent in the items' core metadata. More recent literature in ML and libraries seems less engaged with geographic holdings, but current methods for image and video analysis, alongside corporate investments in geographic data analysis, provide ample reason to consider new experiments with such collections. More recently, the Library of Congress's John Hessler has employed deep learning to extract spatial features from historical maps.

One of the most exciting possibilities of ML work is that it will help expand the features within digital libraries that can even be captured and analyzed as data. Named entities, whether personal, geographic, or otherwise, are already prioritized by catalogs and indices, but there are less obvious patterns that ML can make tractable. In the *America's Public Bible* project, for instance, historian Lincoln Mullen identifies biblical quotations within the Library of Congress's Chronicling America newspaper collection, in order to explore the frequency and context of such quotations in historical

[75] Ryan Cordell, "Machine Learning and Libraries Interview with Amanda Rust," April 29, 2020.

[76] Chen et al., "A Geographic Knowledge Representation System for Multimedia Geospatial Retrieval and Analysis."

newspaper articles (as well as other textual genres reprinted in historical newspapers).[77] Similarly, the Viral Texts project (on which I am a primary investigator) traces the reprinting of texts across millions of newspaper pages.[78]

These and similar projects seek to identify and record very different kinds of metadata from what library catalogs typically represent, tracing not distinct, atomized texts but fragments of texts woven throughout a corpus. As such, they are emblematic of both an opportunity and challenge of ML for enriching library data. The kinds of patterns ML allow researchers to identify may not map easily onto existing metadata structures, a challenge I describe in more detail in 4.1.5. If libraries wish to integrate ML-derived metadata into interfaces for discovery (see the recommendations in 5.4.2), existing infrastructure may not accommodate that integration. However, the prospect of rich, contextual, deeply-linked data is primary driver for much ML work with library collections and may justify such development.

3.2.6. Historical Tabular Data Extraction

In addition to unstructured data (whether textual, visual, auditory, or other), libraries also hold vast archives of structured data in historical media that could be useful for research and teaching, but which are incredibly challenging to digitize with their data structures intact. We might consider a city archive holding historical immigration forms, for example, or a university archive holding paper records from its past students. In an interview, Amanda Rust noted the sheer propensity of "record-style documents" across library collections, including many that are digitized, but only as image or PDF files. Rust pointed to ML as a potentially fruitful path to identifying the data in such collections, extracting and organizing that data into discrete variables, and thus making such collections computationally tractable for research.[79]

Computer scientist Benjamin Lee demonstrated the usefulness of an ML approach to documents in research undertaken for the United States Holocaust Memorial Museum. Lee's project worked with scans of death certificate reference cards produced by the International Tracing Service (ITS), an organization formed to reunite survivors of the Holocaust, and which continues to serve as a reference source for people seeking information about Holocaust victims.[80] Lee identified at least

[77] Lincoln Mullen, "America's Public Bible: Biblical Quotations in U.S. Newspapers," 2016, https://americaspublicbible.org/.

[78] Ryan Cordell and David Smith, "Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines," Digital History Project, The Viral Texts Project, 2020, https://viraltexts.org.

[79] Cordell, "Machine Learning and Libraries Interview with Amanda Rust."

[80] I want to note Lee's well-considered discussion of this project's ethics, given its sensitive subject matter: "it is imperative that any application of these [ML] methods to Holocaust material is done ethically and with the utmost consideration of the victims. I contend that the research presented in this article abides by these stipulations. First, distinction must be drawn between a document in an archive and the person described by the document: this research serves to analyze and classify CNI cards within ITS, not the human beings whom they describe. Second, because card type is an intrinsic property of a CNI card, the classification of CNI cards by type is a task with an a priori answer, unlike subjective tasks related to the document's content, such as sentiment analysis or topic modeling. Finally, though this article includes many figures and frequently quotes statistics describing the efficacy and computational efficiency of this method, the intent is neither to dehumanize nor to aestheticize the content of the archive; rather, these plots and statistics are necessary features of demonstrating the effectiveness of this method in improving access to the unique information and stories preserved within the archive.", Lee, "Machine Learning, Template Matching, and the

four variants of the ITS' death certificate cards, and notes a number of factors that make the digitized cards difficult to research, including the fact that "the only consistent metadata recorded were the names of referenced individuals," the low resolution of the digital images, and the predominance of handwritten information on the cards.[81] Lee employed computer vision and machine learning to classify the cards in this archive. From nearly 40 million cards, Lee's classifier was able to identify more than 300,000 death certificate cards of three distinct kinds.[82] This preliminary ML research enhances researchers' ability to analyze ITS cards while modeling a path for ML methods to help develop more rich, computationally tractable record and document collections.

3.2.7. Visual Data Annotation

In the section on clustering and classification above, I discussed applications of neural networks for mapping relationships among visual materials, such as paintings or photographs, to enable new forms of exploration. In addition to sorting by visual similarity—already potentially exciting for new modes of research—ML methods can help annotate the content of both still and moving image data to make such materials more amenable for browsing and search. The *Civil War Photo Sleuth* project at Virginia Tech University, for example, collates data from the Library of Congress, National Archives, National Portrait Gallery, and other sources to identify soldiers depicted in Civil War card-portraits. This project leverages crowdsourcing to both create data (users can contribute their own photos) and to help verify the results of their ML identifications.

In addition to still images, ML methods can make video data in library collections more tractable for research. Markus Mühling *et al* describe research for automatically annotating more than 2,500 hours of video archives from the German Democratic Republic (GDR). Using a deep convolutional neural network, they developed a content-based video retrieval system that enables researchers to find particular moments, figures, or subjects (e.g. *"Trabant, GDR emblem, military parade, optical industry,* or *community policeman"*) within those videos, as well as text that is displayed in video particular segments (e.g. headline text in newsfeeds or signs in scenes).[83] In this experiment, the neural network tagged 77 concepts and 9 people from their source videos and "The system was extensively tested by an archivist in the everyday work of the German Broadcasting Archive. From this archivist's perspective, it has been shown that concepts with an AP [average precision] score of more than approximately 50% turned out to be very useful in practice. Altogether, 66% of the concepts achieved an AP score of more than 50%."[84] This experiment shows that ML could be quite useful for enriching moving image data, which is a particularly difficult data type to atomize.

In a related article, Taylor Arnold, Lauren Tilton, and Annie Berke demonstrate how face detection and recognition algorithms can enable advanced formal and cultural analysis of moving images. They apply computer vision algorithms toward a "distant viewing" of two mid-twentieth-century

International Tracing Service Digital Archive," pg. 524

[81] Lee, pg. 518.

[82] Lee, pg. 528-529.

[83] Markus Mühling et al., "Content-Based Video Retrieval in Historical Collections of the German Broadcasting Archive," *International Journal on Digital Libraries* 20, no. 2 (June 1, 2019): 167–83, https://doi.org/10.1007/s00799-018-0236-z, pg. 168, 172.

[84] Mühling et al., pg. 178.

US sitcoms, *Bewitched* and *I Dream of Jeannie*, analyzing more than 150 hours of material. Their focus, as they write, is

> on the application of facial recognition algorithms to locate and identify characters with a shot. Details regarding framing, shot blocking, and narrative structure can be inferred directly from information about the characters…Our goal is to build algorithms that are able to extract these semantics automatically from a corpus. Shot semantics provide features that are of a higher complexity relative to shot breaks and color analysis. By incorporating a more nuanced view of the moving images, our analysis not only provides a deeper understanding of the formal decisions made by actors, writers, camera operators, directors, and editors but lays bare those meanings that are not necessarily intended but are still articulated through form and style.[85]

Among other things, this "shot semantics" approach enables the team to analyze the gender dynamics of the two shows by evaluating the prominence of particular characters, based both on the amount of time they are on the screen, but also based on less immediately apparent dynamics, such as which gender of characters typically appear first in the scenes of one show versus another. While there are not yet many ML experiments using moving images in libraries, those such as I have discussed here point to the enormous potential of ML methods for enriching such collections.

3.2.8. Audio Data Annotation

Similarly to moving images, audio data has been particularly challenging to label with granularity and precision. Here, too, ML methods show great promise for making audio data more tractable for exploration and research. Basaran *et al* describe using a convolutional recurrent neural network to estimate the dominant melody in musical recordings, which enables music to be grouped by similarity of melody.[86] Such clustering has significant research implications, allowing researchers to compare and contrast the shape of compositions across collections, time periods, or even genres. As described by the *Dig That Lick* project team, this approach facilitates "automatic recognition of musical structures and their linkage through metadata to historical and social context."[87] The *Dig That Lick* "Pattern History of Jazz" and Pattern Search websites provide compelling examples of ML-driven exploratory interfaces. Users can search the recurring patterns across the project's audio, seeing and hearing the similarities uncovered by the ML clustering and gaining new insight into the "licks" that define the history of jazz music. As libraries support more ML research, this project provides a useful model for how an interface to ML data can enable meaningful search and browsing for researchers, students, and the public.

[85] Taylor Arnold, Lauren Tilton, and Annie Berke, "Visual Style in Two Network Era Sitcoms," *Journal of Cultural Analytics*, 2019, https://doi.org/10.22148/16.043, pg. 5.

[86] Dogac Basaran, Slim Essid, and Geoffroy Peeters, "MAIN MELODY EXTRACTION WITH SOURCE-FILTER NMF AND CRNN," in *19th International Society for Music Information Retreival* (Paris, France, 2018), https://hal.archives-ouvertes.fr/hal-02019103.

[87] "Dig That Lick," 2019, http://dig-that-lick.eecs.qmul.ac.uk/Dig%20That%20Lick_About.html.

3.2.9. Linking Collections

This report will not delve into the literature around linked open data, which is far deeper than I can possibly do justice. However, I do want to note the potential for ML to help identify linked data across collections or even institutions, in particular for automatically mapping metadata. In an interview, Amanda Rust noted how difficult it currently is to link data across small, local repositories.[88] Organizations such as the Digital Public Library of America and Europeana have undertaken enormous projects linking metadata across collections, but doing so requires significant handwork, and consortial efforts are limited. The *The Atlas of Digitised Newspapers and Metadata,* for instance, reports on the *Oceanic Exchanges* project's work to conduct computational analyses across national digital newspaper collections.[89] Researchers are keen to work across collections, and ML could potentially assist in enabling such work.

3.3. Library Administration and Outreach

In addition to aiding discovery, a subset of the ML and libraries literature focuses on the uses of machine learning for administrative and outreach tasks, such as collection management or better understanding of patrons. While some of these practical applications are likely worth further exploration, in 2.2 I caution against patron-focused efforts that move into grey territory in terms of participant consent and strongly advise libraries to be proactive in their responses to these concerns. There are a range of institutionally-focused applications of ML that are worth outlining this report, however, with appropriate caveats.

3.3.1. Collection Management

One task all libraries face is evaluating and weeding collections, either due to space or budget constraints. A study from Wesleyan University found that a classifier trained on librarians' priorities for weeding (e.g. age, number of times checked out, recency of last circulation, number of copies in peer libraries) was able to reliably suggest titles to be removed from circulation, with "statistically significant agreement between human decisions and automated classifier predictions." Importantly, the model suggested by this study does not entirely automate the process of weeding, and favors precision over recall, as "mistakenly discarding an item has more impact than mistakenly keeping it."[90]

While focused on weeding, this study has implications for the other areas outlined in this report, in that it integrates the automation of an ML assessment with the judgement of human experts. The ML classification suggests the titles most likely to need weeding, but these suggestions are then reviewed by library staff before action is taken. This application of ML speeds a labor-intensive

[88] Cordell, "Machine Learning and Libraries Interview with Amanda Rust."

[89] M. H. Beals and Emily Bell, "The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges" (Loughborough, 2020), https://www.digitisednewspapers.net/.

[90] Kiri L. Wagstaff and Geoffrey Z. Liu, "Automated Classification to Improve the Efficiency of Weeding Library Collections," *The Journal of Academic Librarianship* 44, no. 2 (March 1, 2018): 238–47, https://doi.org/10.1016/j.acalib.2018.02.001, pg. 245.

task, but the workflow explicitly incorporates the trained judgment of library professionals. These judgments can in turn be integrated into subsequent training of the weeding algorithm. As libraries establish ML workflows, they should be self-consciously structuring similarly iterative dialogues between ML processes and human expertise.

3.3.2. Preservation and Conservation

Though far less written about than other topics, the Library of Congress's Abigail Potter also points to emerging projects seeking to employ ML to identify collection items in need of "preservation or conservation action." In one case, for instance, a computer vision project reviews photographs of library stacks and seeks to identify visible signs of damage or wear that might direct librarians' attention to areas of need. In concert with digitization, we can imagine a range of ML efforts aimed at collection triage and preservation prioritization.[91]

3.3.3. ML Literacy Education

In 5.5.1 I argue that libraries' existing role as sites for information literacy education should extend to ML literacy. University, school, and public libraries already undertake significant work helping patrons understand the reliability of research sources, and these duties will magnify as AI/ML accelerates the production of unreliable information online and increasingly determines real-life outcomes for patrons, as described in Section 2. Just as existing library pedagogy focuses on helping students and patrons evaluate the credibility, point-of-view, perceived audience, and context of research materials, ML literacy pedagogy would emphasize the situated-ness of ML training data and experiments, including the biases or oversights that influence the outcomes of academic, economic, and governmental ML processes. This pedagogical mission also complements libraries' existing commitments to privacy, helping patrons become more active agents against growing systems of surveillance and algorithmic injustice. By centering themselves in communicating with the public about ML, libraries can take a leadership role in one of the central cultural debates of the twenty-first century.

3.3.4. Supporting Patron ML Experiments

In 5.2.1 I argue "digitization and access remain the biggest challenges to responsible and representative ML experiments using library collections." In brief, creating more machine-actionable collections and making them easily accessible (whether through bulk downloads or APIs) leads to more experimentation outside of the library itself, among computational users in digital humanities, social science, and computer science. A central role for the library in ML, then, will be the continued creation and management of digital collections, as well as continued dialogue with researchers who make use of those collections. In this way, a library's role in ML may be much wider than the projects underway within it. As noted in the introduction to this report, the library can

[91] Abigail Potter, "Report Draft Comments," May 15, 2020.

serve as a unique site for interdisciplinary dialogue and collaboration around collections, but ideally the collaborations that result for such dialogue will flourish within and beyond the confines of any particular institutions. Here too I would point to the example of LC Labs, who both foster ML projects, such as the frequently-cited Newspaper Navigator, but also provide their data to a broader user community in order to foster follow-on projects and experiments.

3.4. Creative and Activist Interventions

While this report focuses primarily on research outputs of ML, one important thread from the ML + Libraries meeting and the broader literature is ML's potential for expressive, artistic, and activist projects. Such work can support a number of the most urgent goals expressed in this report, for example by prompting people to more carefully consider how representation and bias influence the outcomes of ML in corporate, governmental, and other contexts. In Joy Buolamwini's project, "The Coded Gaze," she demonstrates how wearing a white mask makes her more legible to facial recognition software in order to expose the racial biases that informed the algorithm's creation. The project dramatizes the "exclusion from seemingly neutral machines programmed with algorithms" for people of color.[92] This important work contributed to IBM's recent decision to divest from developing or selling facial recognition software due to its racial biases.[93]

Buolamwini founded the Algorithmic Justice League to educate people about these issues and advocate for just technologies, and their work blends art, research, and advocacy. Work such as that undertaken by the Algorithmic Justice League should be central to libraries' strategies for both managing bias (see 2.1), by confronting it rather than ignoring it, as well as for building patrons' ML literacy (3.3.3). I would also point to Kate Crawford and Trevor Paglen's "ImageNet Roulette," in which participants could upload a photograph, which would then be assigned tags based on the ImageNet library's annotations:

> When a user uploads a picture, the application first runs a face detector to locate any faces. If it finds any, it sends them to the Caffe model for classification. The application then returns the original images with a bounding box showing the detected face and the label the classifier has assigned to the image. If no faces are detected, the application sends the entire scene to the Caffe model and returns an image with a label in the upper left corner.[94]

While some of the returned classifications are innocuous, funny, or just strange, others are offensive, and explicitly demonstrate the biases in the training data underlying ImageNet. I delve into ImageNet's problems more in 4.1.3, but for now I point to "ImageNet Roulette" as an ML project that explicitly seeks to build public literacy about the potential ramifications of training data biases for broader ML implementation. As with "The Coded Gaze," this project explains formerly unexplained AI and brings public attention to an urgent problem. As libraries contemplate ML projects,

[92] Joy Buolamwini, "The Coded Gaze," AJL -ALGORITHMIC JUSTICE LEAGUE, November 6, 2016, https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148.

[93] Peters, "IBM Will No Longer Offer, Develop, or Research Facial Recognition Technology."

[94] Kate Crawford and Trevor Paglen, "Excavating AI: The Politics of Training Sets for Machine Learning," -, September 19, 2019, https://www.excavating.ai.

such projects can serve as models for explainable ML that interrogates and confronts bias rather than hiding it behind an interface or visualization.

In addition to activism, artists and scholars are increasingly experimenting with ML methods for generative, aesthetic creations. The screenplay for the short film *Sun Spring*, a collaboration between filmmaker Oscar Sharp and (self-described) creative technologist Ross Godwin, was written by an ML program trained on hundreds of sci-fi movie and television scripts.[95] Hannah Davis is a "generative composer" who creates projects like *TransProse*, which "finds different features throughout a work of text, and programmatically creates music with a similar emotional tone and underlying structure."[96] I would also point to Janelle Shane's fantastic blog *AI Weirdness*—soon to be a book—which documents "experiments [that] have included computer programs that try to invent human things like recipes, paint colors, cat names, and candy heart messages."[97] Most of these projects are as much about interrogating the possibilities and limitations of their underlying technologies as they are about the aesthetic objects created, but as such these experiments are intriguing models of explainable ML, in which readers (or hearers or users) come to better understand how ML works through engagement with the sometimes strange, sometimes delightful, and sometimes unsettling artifacts it produces.

For libraries, these creative engagements should not simply be an afterthought. At the very least, libraries should strive to make digital collections available for generative and transformative use. As I will describe in 5.2, access to data for ML must be the core of any library's ML strategy. Beyond data access, however, libraries should consider the ways generative and creative ML experiments might prompt greater public engagement, both with their collections and with ML technologies. Could an ML-based recipe generator prompt greater engagement with a collection of historical recipes, for instance, or might an auto-tagger for patrons' photos help them better understand how ML was used to classify a collection of artwork? There is ample opportunity for collaboration between libraries and data artists to serve core missions toward public scholarship and information literacy.

4. Common Challenges for ML + Libraries

While there is much excitement about the possibilities for ML in libraries, as described in Section 3, that excitement can be tempered by both practical and intellectual hurdles. Participants in the Machine Learning + Libraries Summit, as well as those interviewed during the preparation of this report, outlined a number of shared challenges for conceiving, planning, and implementing ML experiments. Meeting these challenges will require concerted efforts between institutions, as well as collaboration among librarians, academic researchers, and technologists.

[95] *Sunspring*, 2016, https://www.youtube.com/watch?v=LY7x2Ihqjmc.
[96] Hannah Davis, "TransProse," 2013, http://www.musicfromtext.com/.
[97] Janelle Shane, "AI Weirdness," 2020, https://aiweirdness.com.

4.1. Data for ML

There are several distinct types of data essential to all ML projects. Each must be considered and each presents unique challenges for ML work in libraries. Some projects have more stages and include more data types, but all ML projects in libraries should include:

1. **Training data** in a particular domain, in which the elements the ML project hopes to analyze (e.g. genre, visual content, geographic features) have been annotated by human beings (e.g. through tagging, markup, or similar). ML models learn from the examples provided in the training dataset.
2. **Validation data** (sometimes called "test data," and there is some confusion in the literature between these terms), which is similar to training data but not used to train the model. Once a model is trained, it is applied to the test data to see if it fits as expected.
3. **Unannotated data** in the chosen domain to which the model, once trained and validated, will be applied.
4. **ML-annotated data**, i.e. the new data that results from the ML process.

The challenges I discuss below will pertain to some or all of these data types. It is essential to understand, however, that the central challenges of ML in libraries are challenges of dataset construction and contextualization. Fortunately, these challenges are familiar to librarians. While ML might seem technically intimidating, at base a robust ML strategy will draw on libraries' core strengths of dataset construction and description.

4.1.1. Machine-actionable Data

Despite the technical, infrastructural, and social problems outlined in the rest of this section, basic digitization and access remain the single greatest obstacles to more and more effective ML research in libraries. Much of the meta-narrative about the kinds of research ML methods might enable presumes a fundamental access to digital data, a presumption which does not always align with reality. While the past decades have seen significant efforts toward digitization and digital curation, digital libraries represent only a fraction of the materials held in physical libraries and archives. Moreover, as described in 3.1, the collection and digitization practices for digital libraries have been "colonialist, racist, hetero- and gendernormative, and supremacist in other structural and systemic ways" that require redress.[98] Confronting that reality requires that libraries prioritize more inclusive and comprehensive digitization as a necessary prelude to ML projects.

If broader and more inclusive datasets are not available to researchers, then ML experiments will inevitably reinscribe existing gaps and injustices. Digitization itself, in other words, remains the largest hurdle to ML work in libraries. As Padilla outlines in RO, "To date, much of the work producing machine-actionable collections has not been framed as a core activity. In some cases, the work is simultaneously hyped for its potential to support research and relegated to a corner as an unsustainable boutique operation."[99] Though it may seem too simple, the first priority of

[98] Lorang et al., "Digital Libraries, Intelligent Data Analytics, and Augmented Description," pg. 28.
[99] Padilla, "Responsible Operations," pg. 15.

any ML strategy must be creating machine-actionable datasets. As Laura Nelson said in our interview, "Making machine readable data would do far, far more than any GUI tool you might throw up."[100] Data is more important than a bespoke tool because the latter constitutes a walled garden—potentially interesting, but limited—while a single machine-actionable dataset can spark many experiments, visualizations, interpretations, and arguments, both within the library and from outside researchers.

4.1.2. Ground Truth / Training Datasets

The second most common challenge reported by those engaged in machine learning projects with and for cultural heritage institutions is the relative lack of domain-specific ground truth or training data. While some "standard" training datasets exist in the broader ML literature, these are mostly ill suited to the requirements of libraries and other cultural heritage institutions. To date, most ML projects in libraries have required bespoke data annotation to create sufficient training data. Reproducing this work for every ML project, however, risks wasting both time and labor, and there are ample opportunities for scholars to share and build upon each other's work.

The IDA team notes that the Library of Congress's data was relatively simple to gain access to for their research, but that data was not typically annotated in ways that facilitated the ML experiments they sought to undertake. The IDA report summarizes the training data dilemma of many ML researchers pithily: "as ground truth, *ground truth data proved challenging*." Continuing from this maxim, they claim,

> [W]e had to create ground truth sets ourselves or turn to externally available datasets that provided the type/nature of ground truth information needed. Sometimes, we had to create these ground truth sets because the data did not otherwise exist as verifiable data…In other cases, the nature of the ground truth did not fit with our proposed approach.

The IDA team concludes "that the *bibliographic information and collections-centered metadata previously pursued in libraries is a limited vision of what will be needed for machine learning applications and new areas of research*."[101] To put that another way, the metadata found in library catalogs and databases may not map well onto the annotations needed to pursue an ML project, and neither libraries nor researchers cannot presume that available data will be immediately ready for ML tasks, even if the data is broadly available.

4.1.3. Limits of "Gold Standard" Datasets

Popular libraries for training deep learning or neural net models, such as the ImageNet library widely used for projects using visual data, comprise contemporary and conventionally Western subjects and taxonomies. Kate Crawford and Trevor Paglen's "Excavating AI" models a practice of data archeology to help understand that "every layer of a given training set's architecture is

[100] Cordell, "Machine Learning and Libraries Interview with Laura Nelson."

[101] Lorang et al., "Digital Libraries, Intelligent Data Analytics, and Augmented Description," pg. 29.

infused with politics." Crawford and Paglen focus on ImageNet, a database of "over 14 million labeled images organized into more than 20 thousand categories" that has become ubiquitous as training data for ML research projects. While ImageNet was first announced in 2009, they show its taxonomies are "based on the semantic structure of WordNet, a database of word classifications developed at Princeton University in the 1980s." As such, ImageNet inherits assumptions from WordNet, such as the idea that "male" and "female" are the only two valid types of human body, while other gender identities are present only as outdated and often-offensive terms.

Looking further, Crawford and Paglen find that categories include "racist slurs and misogynistic terms" to describe images of people in the database. They conclude their investigation by emphasizing the power of training sets to shape anything we might learn from ML/AI:

> There is much at stake in the architecture and contents of the training sets used in AI. They can promote or discriminate, approve or reject, render visible or invisible, judge or enforce. And so we need to examine them—because they are already used to examine us—and to have a wider public discussion about their consequences, rather than keeping it within academic corridors. As training sets are increasingly part of our urban, legal, logistical, and commercial infrastructures, they have an important but underexamined role: the power to shape the world in their own images.[102]

The lesson of this kind of investigation for libraries is clear: ML projects should not adopt "standard" training dataset such as ImageNet simply because they are easy to obtain or use. It is essential that projects understand the histories of any existing training data used so that its biases and oversights can, at the least, be addressed. For many ML projects in libraries, existing datasets such as ImageNet are not well suited as ground-truth or training data.

4.1.4. Fitting Data Across Domains/Periods

In addition to the problems of racial, ethnic, class, gender, or sexuality bias outlined above, most widespread datasets in the ML literature also suffer from selection biases that hinder application to many library collections. Standard ML training datasets are focused on (relatively) contemporary materials and are often biased toward dominant racial, linguistic, or cultural groups. In other words, they are not well-fitted to help researchers understand many communities, regions, cultures, or historical periods. When such "out of the box" datasets are used to train a model for cultural heritage work, the resulting analyses will not suffice. In our interview, Kate Zwaard described the Library of Congress's work on the *Speech-To-Text Viewer*, which seeks to transcribe the dialogue from Smithsonian Folkways Recordings. "The accuracy of the transcriptions was quite low," Zwaard described, partly due to the quality of the recordings used, "but also because of regional dialects" in the recordings for which little training data exists.[103]

The LoC's experience in this project demonstrates the kinds of domain mismatch that can hamper effective, just library ML projects. Using an existing language model likely trained on "general American English"—a linguistic category that centers a dialect largely associated with white,

[102] Crawford and Paglen, "Excavating AI."
[103] Cordell, "Machine Learning and Libraries Interview with Kate Zwaard."

middle-class Americans as "standard"—means that the voices of people from marginalized communities will be less tractable to ML and thus to users who might benefit from ML analysis. This kind of domain mismatch is one of the most pressing challenges for libraries delving into ML, as the process of annotating training data is often the most costly element of an ML project, both in terms of salary for those doing the work and in terms of the time required. Libraries and researchers need ground truth and training datasets compiled specifically for cultural heritage work, curated by domain experts.

Moving forward, library ML projects should strive to make not only the data resulting from their analyses, but also their training datasets available post-project. As Benjamin Lee outlined in an interview, ML experiments can benefit enormously from pre-training on near-neighbor data, which can significantly reduce the time and labor of training data annotation. As Lee noted, when bootstrapping on a similar dataset an ML project might need to annotate only a few hundred pages in the project's particular domain rather than a few thousand, or (to put it another way) to only spend a few days annotating training data, rather than a few months.[104] More regularized sharing of training data, in other words, could make ML projects more feasible across the board for libraries.

4.1.5. New Objects of Analysis

In their interview, Lauren Tilton and Taylor Arnold raised an additional, intellectual challenge posed by ML in libraries. Discussing their work on film (discussed in more detail in 3.2.7), they noted that conducing an ML analysis of moving images changes the object of study itself:

> When you start looking at moving images and doing ML on them, your object of study is no longer an entire film but shots or scenes or still images. And that suddenly brings in important questions for archives and libraries. When the record itself changes…that raises huge questions that folks are not grappling with. Not just enriching a particular object but changing the object.[105]

The transformative effects of ML analyses will pose fundamental questions for catalogers, archivists, and information scientists alike about how to document the new objects created through ML work and relate those new objects to the source materials they help illuminate.

4.1.6. Sharing Data

In the same interview, Tilton and Arnold pointed to another major challenge that echoes throughout the literature: while there are sustained and growing calls for ML researchers to share data, "there's no real agreed upon standard for sharing this data or what it should look like. Everyone keeps saying IIIF but that doesn't seem to be happening."[106] Their own work on twentieth-century video materials raises additional questions of intellectual property, and whether the training or output data they generate is sufficiently transformative to be shared openly. These concerns were echoed in my interview with Amanda Rust, who noted that, despite the long conversation around linked

[104] Ryan Cordell, "Machine Learning and Libraries Interview with Benjamin Charles Germain Lee," February 6, 2020.
[105] Ryan Cordell, "Machine Learning and Libraries Interview with Lauren Tilton and Taylor Arnold," March 3, 2020.
[106] Cordell.

open data in the library field and beyond, few libraries have interlinked systems that can make full use of existing data, much less the data that might be derived from ML experiments. Rust pointed to SNAC (Social Networks and Archival Context) as one system just beginning to implement the vision of linked open data.

To effectively share ML data will require some reimagining of library systems, and thoughtfulness about means of sharing that do not require adapting data to legacy metadata systems or vendor-controlled solutions. In my recommendations in Section 5 I will point to projects, such as the LoC's Newspaper Navigator, that seek to model data sharing, but notably most current examples rely on out-of-copyright source materials, meaning the unique challenges raised by twentieth- and twenty-first-century materials remain unsolved. Padilla's RO notes that "Rights assessment at scale presents significant challenges for libraries," which is only compounded by "machine actionable collection use":

> [U]sers seek to analyze large collections (e.g.,thousands, hundreds of thousands, millions of works); make use of content types replete with challenging licenses and terms of use (e.g., A/V materials, social media data); make use of aggregate collections from multiple national sources with competing legal paradigms governing use; and situations arise wherein rights assessment is clearly determined but ethical questions bearing on use remain.[107]

The Library of Congress's Kate Zwaard also acknowledged the challenges of rights assessment, saying, "A huge barrier to putting materials online is rights analysis." Zwaard wondered, however, "if we could use ML to take a first pass at materials that can't be published openly online" in order to "scale up our ability to give access to materials."[108] Tilton and Arnold noted that solving the problem of data access will "require conversations between ML scholars, data scientists, and librarians" and that scholars will look to libraries such as the Library of Congress for guidance on these issues.

4.2. Staff Expertise in ML

Across the board, the greatest practical obstacle to effective ML work in libraries is staff expertise. Few LIS programs provide training in ML methods, which means libraries cannot expect to hire entry-level librarians with the necessary skills to take on ML projects. Trained ML specialists from other sectors command substantial salaries, whether libraries are seeking to hire full-time, in-house expertise or contract expertise for particular projects. Training from within requires substantial investments of time, as existing staff must be freed from other duties—which potentially must be assumed by other staff members—and given sufficient time to cultivate ML proficiencies. Given the overall trend toward constriction in library budgets, it is difficult for many, if not most, libraries to gather the necessary expertise to undertake ML work, however exciting the projects they might imagine. Small and medium-sized institutions in particular struggle with access to expertise.

Given these realities, there is immense pressure for libraries to make use of vendor solutions that

[107] Padilla, "Responsible Operations," pg. 17.
[108] Cordell, "Machine Learning and Libraries Interview with Kate Zwaard."

promise easy access to ML results. These platforms, however, also cede control of collections and processes to vendors in ways that foreclose many of the most innovative possibilities for ML transforming research access. Libraries cannot recreate all of the services that are provided by technology or information companies, and so forms of collaboration and coordination will be required to give libraries the staffing support needed to undertake ML projects.

Conversations about staff expertise tend to focus on technical implementation, but for libraries other ML literacies will be equally necessary. During the ML + Libraries meeting at the LoC, participants noted that many scholars in the humanities and social sciences do not yet fully understand the kinds of questions that ML techniques might help them answer, or how to articulate the work they want to do in language that will appeal to collaborators in the sciences, or even funding bodies. Sociologist Laura Nelson noted the potential of libraries as sites for translation between domains.[109] This will require staff with broad expertise who can help scholars in the humanities or social sciences operationalize their questions in ways that align with the library's collections and potential ML approaches.

Such a staffing model will not require individuals who can do everything, from conceptualization through technical implementation, but instead staff who are literate across domains and can facilitate collaboration. It is this kind of literacy, rather than a narrow technical competency, that might be fostered with relative ease in MLIS programs and professional development training, as I will discuss in my recommendations in 5.5.1 and 5.5.2. Instead of focusing entirely on technical staff, libraries should strive to build cross-functional teams that feature subject experts, collections stewards, data scientists, project managers, and senior-level leaders who will champion ML work.

4.3. Computational Infrastructure for ML

In addition to staffing concerns, participants in the LoC ML + Libraries Meeting cited the development of computational infrastructure as a significant challenge to ML work in their libraries, and a significant portion of the literature directed toward library administrators focuses on questions of hardware and software. Independent scholars, in particular, cited the costs of ML-capable hardware as a significant barrier to their work. Certainly these are primary questions, as ML projects cannot proceed without sufficient computational infrastructure. Fortunately, however, this is an area in which answers are becoming substantially easier.

In the past decade, the infrastructure required for ML projects, even quite computationally demanding ones, has declined precipitously. Mühling *et al* lay out the infrastructure required for their video annotation experiments quite specifically:

> While the execution of models for concept classification and similarity search is really fast on GPUs (only a few milliseconds), training of such deep CNNs [convolutional neural networks] is computationally expensive even on graphics cards. Considering the hardware requirements concerning processing power, GPU memory, main memory and hard disk capacities, we have built a highly optimized system for deep learning similar to the Nvidia DevBox11 to train deep neural network models. The system

[109] Cordell, "Machine Learning and Libraries Interview with Laura Nelson."

> consists of four GeForce GTX Titan X GPUs with 12GB RAM and 3072 CUDA cores at 1000/1075 MHz, an Intel Core i7-5930K CPU with six cores at 3.50GHz, 64 GB of DDR-4 RAM, 8 TB of disk space for large datasets and a 250 GB SSD for fast I/O operations.[110]

In an interview, Benjamin Lee noted that a machine with a few Graphical Processing Units (GPUs) will do most of what is needed for most library ML projects. The Newspaper Navigator project, which Lee led, was able to finetune one of its larger models in 17 hours using a single GPU. The Newspaper Navigator pipeline, once established, was able to process 16,368,041 pages of data in 19 days using "2 g4dn.12xlarge Amazon AWS EC2 instances, each with 48 Intel Cascade Lake vCPUs and 4 NVIDIA T4 GPUs."[111] While these resource requirements are not negligible, they are likely feasible for many library projects. What's more, the Newspaper Navigator models are now published and can be reused for other projects working on newspaper data. As a consequence, Lee recommended that libraries embarking on ML projects buy the necessary hardware, as rented resources (through Amazon Cloud services, for instance) cannot be used for other things when a project is finished and do not pay off in the long run.[112] While it is possible that rented infrastructure could lower the barrier of entry to some organizations, libraries should consider whether the tradeoffs are worthwhile.

4.4. Understanding Full ML Workflow

Participants at the ML + Libraries meeting emphasized the need for libraries to better understand the full ML workflow, from project conception to execution and distribution. The library administrators who were present—as well as some interviewed subsequently—felt they grasped the broad strokes of how ML methods might enhance their collections. In other words, they reported a general understanding of the ML applications outlined in Section 3 of this report, and a desire to explore ML methods with their collections. Many were less certain, however, about the full ML workflow. In other words, how precisely does one bring a collection through digitization and the processing steps required to perform an ML analysis?

Dan Cohen provided an example of this quandary describing a new collection of approximately 1.5 million printed photographs from the *Boston Globe*'s archives (along with 5.7 million negatives), recently acquired by Northeastern University's Archives and Special Collections. Many of these photos, Cohen described, have handwritten annotations on their backs, which would require enormous time and labor to document. It would cost millions of dollars to catalog these materials by hand, Cohen estimated, which makes doing so outside the scope of the library's budget. He wants to explore the use of computer vision APIs to undertake this cataloging, but is not sure all the steps such a workflow will entail.[113] In many ways, this uncertainty about workflow stems from the research-library divide explored throughout this report. Many of the articles cited here

[110] Mühling et al., "Content-Based Video Retrieval in Historical Collections of the German Broadcasting Archive," pg. 173-174.

[111] Benjamin Charles Germain Lee et al., "The Newspaper Navigator Dataset: Extracting and Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America," May 4, 2020, http://arxiv.org/abs/2005.01583.

[112] Cordell, "Machine Learning and Libraries Interview with Benjamin Charles Germain Lee."

[113] Ryan Cordell, "Machine Learning and Libraries Interview with Dan Cohen," February 7, 2020.

outline the workflow for the particular experiments undertaken, but whether these procedures are generalizable to other institutional or research contexts is less clear.

No single workflow or pipeline will pertain across all library ML projects, but certainly more examples are needed, and those examples need to be more accessible to libraries developing ML strategies. In 5.3.2 I point to a few models, recently published, as exemplars of the kinds of guidance needed. More broadly, I recommend prioritization of publications that document process, rather than focusing only on outputs, as such example pipelines will be necessary to broadening the field of ML researchers in libraries and other cultural heritage institutions.

4.5. Integrating ML Research Data Into Library Systems and Infrastructure

Surveying a range of scholarship and projects about machine learning in libraries points to a phenomenon we might name as the "perpetual future tense" of the work's impact. Projects are designed in order to aid browsing or enhance discoverability in collections, say, or to enhance the metadata in catalogs, and data is collected that begins to meet these goals. Looking forward, the literature points to a time when ML will be incorporated into library digitization workflows and help structure core discovery and research experiences. However, at present the results of ML research are rarely ingested or incorporated into the interfaces whereby most users and researchers encounter collections, such that the contributions of ML to discoverability remain largely theoretical rather than practical.

Within libraries one finds significant anxiety about the reliability of ML-derived data. Librarians and researchers alike reckon throughout the field's literature with the consequences of information created through probabilistic means. As Kate Zwaard acknowledged in our interview, "Figuring out how to display and contextualize the results" of ML experiments "is a hard question," in part because the "people doing experimentation in [ML] are often working separately from the people in charge of" library systems.[114] How can the reliability of ML data and metadata be assessed, and how can probabilistic information be integrated with human-created information, or integrated into systems designed around hand-assigned categories, tags, summaries, and so forth? To phrase this central question in another way, the ML and libraries field must develop means to bridge a world that prioritizes expert data and metadata, created slowly, and a set of methods that generate useful but flawed data and metadata, more quickly and at a larger scale.

Mia Ridge summarizes this challenge usefully, thinking about "how staff traditionally invested with the authority to talk about collections (curators, cataloguers) would feel about machines taking on some of that work." Ridge points the way forward imagining "we've broadly moved past that at the library if we can assume that we'd work within systems that can distinguish between 'gold standard' records created by trained staff and those created by software (with crowdsourced data somewhere in between, depending on the project)."[115] ML derived data will likely never meet the gold standard for reliability, but could nonetheless enhance discoverability in some of the ways researchers have promised for decades.

[114] Cordell, "Machine Learning and Libraries Interview with Kate Zwaard."

[115] Mia Ridge, "Museums + AI, New York Workshop Notes," September 18, 2019, http://www.openobjects.org.uk/2019/09/museums-ai-new-york-workshop-notes/.

As a starting point toward addressing this concern, we must note many important digital libraries are currently built atop probabilistic (meta)data, most prominently the OCR-derived text data that underlies the search interfaces of collections such as the Library of Congress's Chronicling America newspapers. OCR has become such a normalized stage in digitization workflows that the probabilistic underpinnings of the technology are rarely considered or highlighted in digital library interfaces. Historical and multilingual OCR remain significant challenges in their own right, as outlined in a previous report authored by myself and David A. Smith, and the text data produced by OCR also likely falls short of any gold standard.[116] Nevertheless, despite its shortcomings OCR has proved immensely valuable for enabling full-text search of large-scale collections, and in fact could be credited with a wholesale transformation of research and access methods for a huge number of scholars, students, and members of the public.[117]

ML methods could result in similar large-scale transformations, but more must be done to integrate promising ML-derived data back into core discovery systems. Amanda Rust noted that libraries' reliance on vendors for cataloging systems makes this integration more challenging, as the process for incorporating ML-derived data is not always straightforward for systems the library does not directly maintain.[118] In addition, ML data raises anxieties about reliability. As Dan Cohen laid out in an interview, "Here we have a world that wants expert metadata and here we have a world that generates 'pretty good' metadata, at like 80%. How do we bridge the two?"[119] Librarians need confidence that ML processes are reliable enough to guide discovery and research, and ML processes must be communicated to users in ways that are responsible without being distracting.

In 5.4.2 I recommend focused development of pilot ML-integrated library interfaces to begin bridging the collections-research gap. Bridging this gap with confidence will require systems in which algorithmic processes and human judgment interact iteratively, so that library experts can be confident that the ML-derived data provided to their patrons is both useful and responsible. To reiterate, the integration of OCR into library systems provides a model, particularly new integrations that allow users to benefit from OCR while being made aware of its limitations. Recently more digital libraries have begun to forefront their OCR, even occasionally reporting confidence scores within the primary textual interface or enabling crowdsourced OCR correction, as does Australia's Trove Digital Newspaper Archive. Development of similar interfaces around other ML data will have a knock-on effect of helping center the library in efforts toward ML literacy.

5. Recommendations

This recommendations section is structured first by a broad goal and then by timeframe. In general, the goals in each section are listed in temporal order, from the most immediately feasible to the long-term. Some recommendations address particular institutions (e.g. the Library of Congress) or institution types, while others are broadly applicable. While I have attempted a fair survey of the field, I am also sure this list will reflect my own priorities and biases as an academic researcher

[116] Smith and Cordell, "A Research Agenda for Historical and Multilingual Optical Character Recognition."
[117] Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago."
[118] Cordell, "Machine Learning and Libraries Interview with Amanda Rust."
[119] Cordell, "Machine Learning and Libraries Interview with Dan Cohen."

who works with particular, field-specific library collections. I hope those working in other positions, fields, or institutional contexts will see these items as invitations to dialogue rather than declarations of boundaries.

5.1. Cultivate Responsible ML in Libraries

Questions of representation, responsibility, and justice must be central to ML work in libraries. Models for structuring such work are emerging from scholarly, non-profit, governmental, and even some corporate sectors. Libraries considering ML projects (or those considering ML projects using library resources or data) must first consider how this work will not simply mitigate bias, but work toward fair representation and algorithmic justice. As the team who developed the "Algorithmic Equity Toolkit" argue, "many meaningful interventions toward equitable algorithmic systems are non-technical," and require instead the kinds of "community organizing" and "public engagement" at the heart of libraries' missions.[120] The scholarship discussed in Section 2 of this report provides only a starting point into the literature around algorithmic accountability, but any organization taking on ML for the first time must first dive deeply into this scholarship in order to undertake ML work responsibly. To sum up this section, I would simply amplify Padilla's overarching insistence in RO that a framework of responsible operations must undergird all ML work in libraries.

5.1.1. Commit to Community-Driven Pledges

As libraries work to develop their own statements of ML values, they can look to existing models and consider committing to community-driven pledges that align with their own values and goals. Initiatives such as the Algorithmic Justice League have generated a range of outcomes, including pledges for community responsibility and accountability.[121] One example of this is the Safe Face Pledge, through which "organizations…make public commitments towards mitigating the abuse of facial analysis technology." Unlike a statement of values, which is primarily inward-looking, such pledges emphasize shared responsibility and accountability among organizations and publics. Libraries considering ML projects should consider signing existing community-driven pledges, or coming together to draft (and sign) similar commitments tailored to the cultural heritage sector. Such public commitments serve to amplify organizations' values, as expressed in SoVs, and hold them up to public scrutiny.

Libraries undertaking ML research might also consider signing onto the "Toronto Declaration," which was released by Amnesty International and Access Now in May 2018 and focuses on "Protecting the right to equality and non-discrimination in machine learning systems." In its preamble, the "Toronto Declaration" recommends:

> States and private sector actors should promote the development and use of machine learning and related technologies where they help people exercise and enjoy their human rights. For example, in healthcare, machine learning systems could bring ad-

[120] Katell et al., "Toward Situated Interventions for Algorithmic Equity."

[121] Rachel Thomas, "16 Things You Can Do to Make Tech More Ethical, Part 2," fast.ai, April 25, 2019, https://www.fast.ai/2019/04/25/ethics-action-2/, @algorithmicjusticeleague2018.

> vances in diagnostics and treatments, while potentially making healthcare services more widely available and accessible. In relation to machine learning and artificial intelligence systems more broadly, states should promote the positive right to the enjoyment of developments in science and technology as an affirmation of economic, social and cultural rights.[122]

Given their public missions and, for many, public funding, libraries might pay particular attention to Section 7, "State use of machine learning systems," which focuses on transparency, accountability, and systems of redress for any ML systems that impact citizens. For academic libraries, the same affordances should apply to students, researchers, and other patrons.

Existing community pledges will not fully address the particular needs of libraries, whose missions are distinct from the technology and corporate sectors typically addressed by current initiatives. Nevertheless these documents provide useful grounds for conversation, both locally and across institutions, and could provide initial stepping-stones for libraries seeking to commit to responsible ML work. Where libraries' goals and values align with existing efforts, they should consider co-signing, which serves the additional purpose of making libraries visible within the broader cultural conversation about ML.

5.1.2. Convene Working Group to Write Model Statement of Values for Library ML

The research and library communities have alike called again and again for libraries to develop statements of values for machine learning research that engages responsibly and justly with the communities whose data makes that research possible, and those that make use of the research's outcomes. In their recommendations to ML researchers, Jo and Gebru look to the archival community's guiding mission statements as models for ML researchers, as "a public Mission Statement forces researchers to reckon with their data composition by guiding the data collection process."[123] Despite these models within the library community, calls for SoVs related to ML work are so frequent that while in reality urgent, they risk become routinized gestures. This report seeks to do more than simply repeat them, but instead recommends that the Library of Congress, in partnership with other cultural heritage organizations, take the lead in crafting such a statement as a model for the larger scholarly community.

Padilla recommends in *Responsible Operations* that libraries should:

> Hold symposia focused on surfacing historic and contemporary approaches to managing bias with an explicit social and technical focus. The symposium should gather contributions from individuals working across library organizations and focus critical attention on the challenges libraries faced in managing bias while adopting technologies like computation, the internet, and currently with data science, machine learning, and AI. Findings and potential next steps should be published openly.[124]

[122] Amnesty International and Access Now, "The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems," May 16, 2018, https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf. pg. 2.

[123] Jo and Gebru, "Lessons from Archives," pg. 5.

[124] Padilla, "Responsible Operations," pg. 10.

I would pair this recommendation with Lorang *et al*'s from IDA, that the Library of Congress "Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage."[125] Together, these two ideas suggest a specific path forward that can provide a model SoV for library work in ML more broadly.

Within six months of this report's publication, the Library of Congress should convene a machine learning and libraries working group with an explicit commission to draft a Statement of Values for Machine Learning in Libraries. This commission should be limited, with a set period of time set aside for the group to meet, plan, and draft. In order to work efficaciously, the convened LoC group can look to existing models, such as the "Toronto Declaration" discussed above, or the Association for Computing Machinery's "Code of Ethics and Professional Conduct."[126] The latter has a broader mandate than this working group, discussing ML among a range of other technologies, while the former has a narrower mandate.

The authors of the "Toronto Declaration" note that it focuses on "the right to equality and non-discrimination," while ML systems influence many other aspects of human rights they cannot fully address. Such a declaration, then, will not be entirely sufficient to ML work in libraries, but provides a starting point on which the LoC panel can draw and build. If focused specifically on production, an intensive writing workshop of a few days' length could make significant progress toward a model SoV for ML in libraries. Once drafted, this model SoV for ML in libraries should be published on the LoC website and promoted among the Library's networks, including other libraries, educational institutions, governmental organizations, and non-profits. Libraries should be invited to join the SoV as co-signatories, or to adapt to their own local circumstances or projects.

5.1.3. Adapt Model Statement of Values

Once the LoC working group publishes its model SoV for library ML research, libraries undertaking ML projects should closely review the model and adapt it to local circumstances, communities, and institutional priorities. While the model advocated in 5.1.2 will be important to the wider community, it will not be able to address the distinct needs of all libraries or their patrons, and so local adaptation will be necessary. Before embarking on ML work, then, libraries must consider existing models, such as an SoV developed for the Library of Congress, and adapt it to their own context and project(s). This work should be pro-active rather than reactive, and I would recommend early development, such that the SoV can help guide subsequent project proposals, funding applications, or discussions around collaboration.

5.1.4. Implement Algorithmic Impact Assessments and Checklists to Guide Development of Justice-Oriented ML Projects

Libraries, universities, and cultural heritage institutions must prioritize responsibility—to their patrons, students, and communities—in any use of ML methods. This will require those institutions

[125] "Image Analysis for Archival Discovery (Aida)," pg. 31.
[126] Don Gotterbarn et al., "ACM Code of Ethics and Professional Conduct," 2018, 28.

to take seriously ideas of algorithmic accountability and institute plans for assessing current and future ML projects. Caplan *et al*'s report on algorithmic accountability focuses on corporate applications but includes principles that are more broadly applicable. Algorithmic accountability, they argue, "ultimately refers to the assignment of responsibility for how an algorithm is created and its impact on society; if harm occurs, accountable systems include a mechanism for redress" while "accountability must be grounded in enforceable policies that begin with auditing in pre- and post-marketing trials as well as standardized assessments for any potential harms."[127]

While libraries do not typically conceive of projects in terms of "pre- and post-marketing," the larger point outlined in this report applies. Researchers involved in ML projects must have ethical standards that can be assessed iteratively throughout the project's progress, and procedures for redressing harms or inequities if they emerge as a result, either during or after the experiment. Such assessment processes will be especially vital if libraries hope to incorporate ML-derived data or metadata into their catalogs or discovery interfaces, and so must be developed early and reassessed regularly as libraries delve into ML techniques.

In their 2018 report "Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability," the AI Now Institute recommends that organizations, and particularly public organizations, "consider incorporating [Algorithmic Impact Assessments (AIAs)] into the processes they already use to procure automated decision systems or any existing pre-acquisition assessment processes the agency already undertakes." AIAs, they continue,

> are explicitly designed to engage public agencies and the people they serve on these areas of concern through the various notice, comment, review, and due process elements. This allows a wide range of individuals, communities, researchers, and policymakers to participate in accountability efforts.

In particular, an AIA should be undertaken before an ML process is adopted, giving "both the agency and the public the opportunity to evaluate the adoption of an automated decision system before the agency has committed to its use."[128] This report goes on to outline the necessary components of an AIA, including defining the scope of the automated system; notifying the public about the system and its potential effects; assessing the effectiveness, biases, and effects of the system's implementation; and giving appropriate access to a range of reviewers.

Similarly, the ACLU of Washington published the *Algorithmic Equity Toolkit* (AEKit) in April 2020.[129] The AEKit "is a collection of four components designed to identify surveillance and decision-making technologies used by governments; make sense of how those technologies work; and pose questions about their impacts, effectiveness, and oversight." The AEKit's flowchart and system map provides a straightforward, beginner-friendly overview of what is and is not an AI or surveillance tool. I want to especially highlight the AEKit's Fill-In-the-Blank and Questionnaire tools, however, which ask a series of questions about a given software or system to help users think through both intended and

[127] Robyn Caplan, Lauren Hanson, and Jeanna Matthews, "Algorithmic Accountability: A Primer" (Data & Society), accessed January 28, 2020, https://datasociety.net/output/algorithmic-accountability-a-primer/, pg. 22-23.

[128] Dillon Reisman et al., "Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability" (AI Now Institute, April 2018), pg. 7-8.

[129] Critical Platform Studies Group, "Algorithmic Equity Toolkit" (ACLU of Washington, April 2020), https://www.aclu-wa.org/AEKit.

unintended consequences before implementation. The questionnaire, for example, explicitly asks users to consider "Who gathered that data [this system is using], with what tools, and for what purposes?" and concludes asking, even "If the system works without errors, does it still perpetuate injustice?"[130] The materials in the ACLU's toolkits are especially useful, then, for thinking through both intended and unintended consequences of ML systems.

While the AI Now Institute and ACLU of Washington documents both focus primarily on systems with immediate, material impacts on users, their recommendations could be adapted to the library and cultural heritage sectors. In particular, libraries can establish clear processes for evaluating the implementation of ML initiatives; educating their communities of researchers and users; and seeking feedback from those communities before, during, and after the implementation of ML processes for library collections or procedures. Following the recommendations in the book *Ethics and Data Science*, Rachel Thomas argues that checklists "can…help those working in tech make fewer ethical mistakes."[131] If, as outlined in RO, the primary directive for ML projects in libraries is to mitigate rather than avoid bias, a checklist can help researchers and developers ensure they consider each of the essential ethical facets of a new project in turn, and that nothing is missed due to oversight alone. In Section 6 of this report, "Questions to Guide Structuring an ML Project," I outline a template for such a checklist focused on the library and cultural heritage sectors.

5.1.5. Develop Toolkits for Explainable ML in Library Systems

One of the primary guiding principles for ethical ML and AI is "explainability." In brief, any decision made as the result of an ML process should be explainable across multiple registers so that a range of users, with differing needs and levels of expertise, should be able to understand. As the researchers behind IBM's "AI Explainability 360 Open Source Toolkit" explain,

> [A] doctor trying to understand an AI diagnosis of a patient may benefit from seeing known similar cases with the same diagnosis; a denied loan applicant will want to understand the main reasons for their rejection and what can be done to reverse the decision; a regulator, on the other hand, will want to understand the behavior of the system as a whole to ensure that it complies with the law; and a developer may want to understand where the model is more or less confident as a means of improving its performance.

As the AINow "Algorithmic Accountability and Policy Toolkit" explains, the fact that ML systems learn from training data presents a significant challenge to understanding their outputs. Because of this fact, the "source code of a system that uses machine learning will not reveal the 'rules' the machine learning model uses to make decisions."[132] Instead, training data can be evaluated and, ideally, the system will be designed to report how that training data was evaluated and weighted in making its decisions.

On the "AI Explainability 360" toolkit's website, users can walk through a demo that simulates an

[130] Critical Platform Studies Group, pg. 7.

[131] Rachel Thomas, "16 Things You Can Do to Make Tech More Ethical, Part 1," fast.ai, April 22, 2019, https://www.fast.ai/2019/04/22/ethics-action-1/.

[132] AI Now, "Algorithmic Accountability Policy Toolkit," October 2018, pg. 5.

ML-derived decision on a home-equity line of credit application, taking the perspective of a data scientist, loan officer, and bank customer in turn. The demo shows how a system could explain the reasons a given application was approved or denied—which factors of the application compared most tellingly to other applicants in the dataset, and which weighed most heavily in the decision—in ways that meet the distinct needs of these three audiences.[133] While this toolkit and its documentation imagine primarily corporate or healthcare applications, it offers a helpful model for cultural heritage institutions using ML to enhance discovery or research systems for their users.

As central nodes between the research, educational, and the public sectors, libraries could play an integral role in helping the wider ML field develop, test, and deploy explainable ML systems. As Michael Ridley notes, "It is concerning that these [ML] innovations are happening outside the field of academic librarianship and with little or no involvement of library expertise."[134] In an interview, Library of Congress Innovator-in-Residence Benjamin Lee identified human-computer interaction (HCI) as an area of particular promise for ML research in libraries. Libraries can draw on their deep expertise in public scholarship and pedagogy to imagine innovative, effective modes of communicating probabilistic data to their users.[135]

When data or metadata in a digital library are derived using ML techniques, those techniques should be explained to users, and ideally explained in different registers for different sets of users: e.g. the domains-specialist researcher vs. the computer scientist vs. the interested member of the public. For example, in a collection of paintings grouped using computer vision, the principle of explainability would ask that the variables for determining similarity—use of color, shape of line, etc.—be described as users navigate that data. For the computer scientist, that explanation might include the formulas used to weight one variable against another, while for the general public that explanation might simply order the variables in descending importance.

Importantly, the bar for explainability varies depending on the project and ML methods used. More normalized techniques like OCR, while not as simple or frictionless as some might imagine, nevertheless are likely to require less explanation than methods employing, say, a convolutional neural network to label the content of images in a collection of artwork. In both cases, libraries should strive to make the algorithmic processes legible to users, but legibility will mean *very* different things in these two cases. Essentially, however, explainability will require that libraries forefront ML processes rather than integrating ML-derived data into existing metadata fields. Content tags assigned to a photo by a neural network should be contextualized as such and accompanied by confidence scores, so that users browsing that collection can benefit from ML processing while understanding the provenance of the metadata they are using.

Explainable ML addresses one of the primary hesitations about applying ML in the library context: bridging a discipline founded on particular, expert curation and a set of methods designed for generalizable classification at scale. If the decisions made by ML are explained, the library's expertise can help explain, refine, amend, or entirely revise those processes to better reflect particular collections or research priorities. Ridley discusses "authorizations" as a primary mode of AI explainability, in which "third parties…provide an assessment or ratification" of the results from

[133] IBM, "AI Explainability 360 Open Source Toolkit," GitHub, 2019, http://aix360.mybluemix.net/.

[134] Michael Ridley, "Explainable Artificial Intelligence (RLI 299, 2019)," *Research Library Issues*, no. 299 (2019): 28–46, https://doi.org/10.29242/rli.299.3, pg. 38.

[135] Cordell, "Machine Learning and Libraries Interview with Benjamin Charles Germain Lee."

an AI/ML process.[136] Such assessments are well suited to library workflows and the expertise of librarians and other domain expertise, and provide a compelling model for blended AI and human processing of library collection data. Much of the literature on AI or ML explainability has focused on corporate, governmental, or scientific applications. Libraries, humanities scholars, and other cultural heritage professionals need a toolkit outlining a theory and praxis of explainability as it relates to historical, cultural, and aesthetic media.

5.1.6. Audit Implementation of ML Ethics Statement(s), Algorithmic Impact Assessments, and Explainable Interfaces

The accountability measures outlined above will only be meaningful if they are regularly implemented and evaluated—they are useless as rhetoric alone, and must be applied in order to ensure ML projects abide by the good intentions they evidence. ML experts recommend regular audits of these policies and procedures. At least initially, such audits must be relatively frequent, as libraries work through initial ML projects and assess their impacts. Specifically, libraries should plan full ML-plan audits:

1. After the completion of a pilot project (see 5.4.1)
2. After two years of ongoing ML experiments
3. After five years and the implementation of a long-term ML strategy

5.1.7. Commit to Honest Reporting

This recommendation is perhaps less tangible than those above, but I wanted to call out an important idea from my interview with Elizabeth Lorang, one of the primary investigators in the IDA project cited throughout this report. In our conversation, Lorang noted, "in our smaller experiments, there would have always been a way to frame those experiments as more successful than they were" while urging other ML researchers to "be mindful and cautious about over-claiming success. The goal should not be success at any costs, or to frame projects as more successful than they are in order to implement them."[137] Particularly in funded research, scholars can feel immense pressure to demonstrate that a proposed approach worked as imagined, even that a project deserves follow-on funding. There can be understandable motives for claiming success beyond simple prestige, if for example staff members depend on a project for employment.

Given their duty of responsible care, however, libraries (and scholars working with library collections) must be especially mindful, as Lorang advocates, of valuing the perceived success of an experiment over its actual value to a collection and community. As the library community continues to build sets of model ML work, examples of failed experiments will prove as useful as successful ones. Though difficult, scholars should seek to report honestly when projects do not meet expectations, and indeed to report in the same detail about "failed" pipelines, data sources, or domain questions as they would about successful ones. To enable such honesty, however, supervisors and

[136] Ridley, "Explainable Artificial Intelligence (RLI 299, 2019)," pg. 35.
[137] Ryan Cordell, "Machine Learning and Libraries Interview with Elizabeth Lorang," February 21, 2020.

funders must commit to acknowledging and rewarding productive failures in ML experimentation, even when it does not result in a product or pipeline that can be widely implemented.

5.2. Increase Access to Data for ML

As described in Section 4.1, access to data is the single greatest practical hurdle to more ML work in libraries. In order to benefit from ML experiments, then, libraries must commit to workflows that make data available for ML analysis and make the results of ML experiments, including process products such as training data, more readily available for potential collaborators. Chris Bourg argues that libraries need to take ML and related processes seriously in their core conceptions of mission and audience:

> I think it will be crucial that we avoid the temptation to continue to serve primarily individual human readers and let the computer scientists worry about how to apply machine learning and AI to vast libraries of resources.
>
> I think we would be wise to start thinking now about machines and algorithms as a new kind of patron — a patron that doesn't replace human patrons, but has some different needs and might require a different set of skills and a different way of thinking about how our resources could be used.[138]

By framing ML as a kind of patron, perhaps, libraries can better cultivate staff expertise, infrastructure, and workflows that will facilitate more routinized and seamless engagements with ML across the library's operations.

5.2.1. Prioritize Access to Machine-Actionable Data

While scholars, administrators, and funders alike focus primarily on the exciting new kinds of analyses and access promised by ML methods, those methods will only ever be as good as the collections to which they can be applied. Digitization and access remain the biggest challenges to responsible and representative ML experiments using library collections. If we hope to see transformative ML projects in libraries, then libraries must re-dedicate themselves to digitizing and providing machine-readable access to those collections, and ensuring what is available to researchers reflects their communities and institutional priorities.

However, it is important to acknowledge that digitization alone will not necessarily facilitate ML experiments. Digitization projects must prioritize the creation of "machine-actionable collections," which "lend themselves to computational use given optimization of form (structured, unstructured), format, integrity (descriptive practices that account for data provenance, representativeness, known absences, modifications), access method (API, bulk download, static directories that can be crawled), and rights (labels, licenses, statements, and principles like the 'CARE Principles

[138] Chris Bourg, "What Happens to Libraries and Librarians When Machines Can Read All the Books?" Feral Librarian, March 17, 2017, https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/.

for Indigenous Data Governance')." In the call for machine-actionable collections, I again amplify Padilla's recommendations in RO:

> In order to move it to a core activity, machine-actionable collection workflows must be oriented toward general as well as specialized user needs. Workflows must be developed alongside, rather than apart from, existing collection workflows. Workflows conceived in this manner will help build consensus around machine-actionable collection descriptive practices, access methods, and optimal collection derivatives.[139]

In short, the core need for cultivating more and more innovative ML experimentation is data that is "getable" and provided in formats that make it easy to work with.

The Library of Congress's "Digital Strategy" demonstrates the core commitments that would facilitate substantial ML work. In addition to a commitment to "continue our aggressive digitization program," the Library also aims to

> prioritize collection readiness by providing more bulk data downloads of content and metadata, improving the usability of the interfaces we publish for programmers, and building more capacity for automated metadata creation, including optical character recognition and speech to text. When the data sets we provide are updated or enhanced, we will make the version information clear to enable reproducibility of research. We will also explore offering tools for non-consumptive data analysis of our content.[140]

Importantly, this strategy commits not only to access, but to documenting the provenance of datasets in ways that will facilitate responsible scholarship. Such commitments also take up Bourg's call to consider ML a library patron with particular needs collections can be designed to meet.

In closing this section, I would amplify Padilla's RO recommendation that digital collection development

> [p]rioritize the creation of machine-actionable collections that speak to the experience of underrepresented communities. Inform this work through collaborations with community groups that have ties to collections, subject experts, and reference to resources produced by efforts like the Digital Library Federation Cultural Assessment Working Group and Northeastern University's *Design for Diversity*. Per community input, decisions to not develop a machine-actionable collection are as positive as decisions to develop a machine-actionable collection.[141]

The goal of digitization should not be to simply make more collections tractable to digital analysis, but to digitize and publish in ways that directly address systemic oversights and advocate for representation and data justice. Doing this well will require working directly with communities and building genuinely diverse teams of researchers—not only in terms of discipline, but also in terms of identity.

[139] Padilla, "Responsible Operations," pg. 15.
[140] Library of Congress, "Digital Strategy," April 26, 2019, https://www.loc.gov/digital-strategy, pg. 3-5.
[141] Padilla, "Responsible Operations," pg. 17.

5.2.2. Build and Share Domain Training Data

Building on the above, I would strongly echo the IDA report's recommendation that the Library of Congress "Focus efforts on developing ground truth sets and benchmarking data and making these easily available."[142] While IDA addresses its recommendations to the Library of Congress specifically, I would amplify their call to focus on building training data sets, which resonates strongly across all the literature, interviews, and meetings surveyed for this report. Rather than conceiving of ML projects as isolated experiments, the community would benefit enormously from standardizing expectations for publishing both training and output data from ML projects.

The Library of Congress's recently-released Newspaper Navigator project can serve as a model for what such publication should look like. Alongside the results of the project's ML experiments, the project team also published its training data and all the code used to obtain its results. That code and, especially, that training data can now benefit a range of future projects working on newspapers, or even similar media such as magazines.[143] There are models in other domains, such as the digital humanities. The *Journal of Cultural Analytics*, for instance, has published a number of ML studies alongside other data-intensive humanities articles. For each article published, *CA* also publishes a dataset description which in turn links directly to a repository of the experiment's data, which other researchers could employ for replication or further development.

As the IDA report outlines,

> Ground truth data and benchmarks will allow researchers—including cultural heritage professionals, computer scientists, and developers—to focus their energies on research, development, and analysis, rather than on creating one-off, niche datasets. The availability of ground truth and benchmarks also create the possibility of more rapid development around particular problem domains.[144]

As described in 4.1.4, shared training data sets will not only help researchers working in the specific domains of particular projects (e.g. eighteenth-century novels or twenty-first century TV sitcoms), but also those working in nearby domains, who could potentially pre-train their training data with "near neighbor" data, thereby reducing the time, labor, and cost of training.

As Padilla argues in RO, "Widening the circle of organizational participation" in ML "could be aided by open sharing of source data and 'gold standard' training data (i.e., training data that reach the highest degrees of accuracy and reliability)." Padilla ties this imperative to the recognized failures of existing training data such as ImageNet (see 4.1.3), and notes that collaborative effort would help create more representative collections:

> Organizations with less representative collections may benefit from developing or being provided the means to combine similar collections (e.g., format, type, topics, periods) across separate organizations in order to produce sufficiently representative datasets. Entities like the Library of Congress and the Smithsonian Institution and/or organizations like Digital Public Library of America (DPLA) and Europeana might aid

[142] "Image Analysis for Archival Discovery (Aida)," pg. 32.
[143] Lee et al., "The Newspaper Navigator Dataset."
[144] "Image Analysis for Archival Discovery (Aida)," pg. 33.

smaller, less well-resourced institutions by facilitating corpora creation and collection classification via crowdsourcing platforms.[145]

Participants at the ML + Libraries event cited crowdsourcing as a particularly attractive venue for creating annotations at scale useful for ML tasks, though libraries should be wary of offloading labor to entirely uncompensated communities. In the next section, I will discuss possibilities for meaningful community participation in ML annotation tasks.

The IDA report recommends the creation of "DocuNet," a database of ML-annotated historical images, to include both "taxonomic and typological metadata"—that is, both data about the content of historical materials and data about the properties of the digital images themselves. Jo and Gebru recommend the development of "data consortia" for ML research, following the models of shared resources such as OCLC or HathiTrust in the library community, which can benefit both large research partners and smaller institutions who gain access to larger collections.[146] Certainly a database or consortial approach to ML training data would be an enormous boon to ML researchers and humanities scholars, who would be able to embark more quickly on ML projects and benefit from prior work. However such a recommendation also introduces new overhead in terms of development, maintenance, and labor.

For the immediate future, then, I would recommend something more modest: sustained development, shared among libraries engaged in ML research, of bibliographies or resource guides to ML training data sets. Rather than seeking to move vast quantities of data, which may be distributed across various institutional or corporate repositories, this more modest goal is simply to gather pointers to those repositories so that researchers undertaking a new ML project can easily identify potentially useful training data sets. Such work could be spearheaded by institutions such as the Library of Congress, organizations such as the Digital Library Federation, or consortia such as the Digital Public Library of America, and should be shared publicly for adoption and adaptation. Ultimately, however, this approach capitalizes on the core competencies of libraries rather than introducing a new service or technical platform.

5.2.3. Enhance Opportunities for Community Participation

Across the literature, one finds excitement about crowdsourcing as a means to annotate training data, verify results of ML experiments, and contribute to ongoing, iterative projects that enhance collections. As discussed in 3.1, currently most library crowdsourcing projects are not primarily designed for ML applications, but instead as means of public outreach that showcase collections. As Meghan Ferriter notes, however, this "state of play is changing, as organizations commit to outside support (for example from the Zooniverse or FromThePage), learn more from outcomes over the last 10+ years of large-scale (web-based) crowdsourcing projects." As an example, Ferriter points to her own "AHRC grant with Mia Ridge and Sam Blickhan that seeks to move the discussion toward the next phase of digitally enabled participation (and in person participation) that foregrounds ethical considerations, data privacy, data utility, clearly communicated goals, and

[145] Padilla, "Responsible Operations," pg. 14.
[146] Jo and Gebru, "Lessons from Archives," pg. 7.

extensible workflows."[147]

As Ferriter's comments about ethical considerations and privacy remind us, libraries should be wary of exploiting free labor and likewise skeptical of "if we build it, they will come" models of crowdsourcing, which assume public interest and investment that does not always manifest. Given the resource limitations discussed throughout this report, libraries certainly want to avoid investing significantly in building infrastructure for crowd contributions before knowing whether there is a crowd interested and available to help with data annotation. Where possible or necessary, libraries might build on existing infrastructure such as Zooniverse, FromThePage, or Transkribus to support crowdsourcing efforts.

Certainly there are materials for which committed communities would be eager to assist in building training data, particularly when doing so will enable new modes of research on understudied collections or materials. These communities may be scholarly, focused on areas of academic interest where ML might make previously intractable collections amenable to computational exploration and evaluation. In addition, there are cultural communities that may be especially interested in helping make their materials more visible through community efforts. Jo and Gebru argue that libraries' existing community archiving projects offer models of potential ML data collection that is decentralized and "also enables minority groups to consent to and define their own categorization."[148]

Munyaradzi and Suleman, for example, describe a project that employed volunteers to transcribe historical notebooks, including text in the |Xam and !Kun languages. This experiment employed ML methods to compare and evaluate volunteers' transcription, and found a high degree of reliability from their contributions.[149] In our interview, Zwaard emphasized that ML "affords a way for crowdsourced contributions to have more power. There's more value from small units of labor than with other kinds of projects," such that motivated effort by scholarly or other communities can potentially have a much larger impact than in other kinds of crowdsourcing projects.[150] While crowdsourcing efforts are often conceived as long-term, passive efforts—in that a platform is created and contributions solicited over time—there are domains where significant improvements could be realized for domain training data through intense, sustained effort on the part of invested communities of experts. In the IDA report, Lorang *et al* recommend that the Library of Congress "[s]ponsor challenges for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas."[151]

As more projects publish training data and pipelines, there will be a growing well of pre-training datasets, as described in 5.2.2. Projects in nearby domains will require significantly less training data to mount ML experiments, meaning that focused challenges could generate exciting new work with relatively modest commitments of time and resources. We see this kind of additive benefit in

[147] Ferriter, "Report Draft Comments."

[148] Jo and Gebru, "Lessons from Archives," pg. 6.

[149] Ngoni Munyaradzi and Hussein Suleman, "A System for High Quality Crowdsourced Indigenous Language Transcription," *International Journal on Digital Libraries* 14, no. 3 (August 1, 2014): 117–25, https://doi.org/10.1007/s00799-014-0112-4.

[150] Cordell, "Machine Learning and Libraries Interview with Kate Zwaard."

[151] "Image Analysis for Archival Discovery (Aida)," pg. 34.

the Library of Congress's Newspaper Navigator project, which was able to use crowdsourced data from Beyond Words as part of its training set. Beyond Words was developed by LoC Innovator-in-Residence Tong Wang, whose work then contributed to the subsequent Newspaper Navigator project developed by Innovator-in-Residence Benjamin Lee. Subsequently, the data Newspaper Navigator generates about twentieth-century newspapers could provide useful pre-training data for a challenge around a nearby domain, such as earlier newspaper data or contemporaneous magazine data.

Libraries planning ML projects should consider from the outset what communities—academic or otherwise—might be mobilized to annotate training data and plan events to train those communities or even work together. The launch data jam for the Newspaper Navigator project, held on May 7, 2020, is an excellent example of such an event, inviting librarians, humanists, computer scientists, data scientists, and artists to engage with the project's data and consider new research possibilities.[152] Funders and libraries with more resources should take up IDA's recommendation and fund more sustained training data challenges, mobilizing communities around collections for which focused efforts could substantially advance the potential for ML exploration and analysis.

5.2.4. Encourage Creative Reuse of Collections

In 3.4, I outlined the important interventional and educational work undertaken by creative and activist ML projects. While access to machine-actionable data is often conceived with researchers as a primary audience, creative reuse, remixing, and experimentation will be central to public engagement with and understanding of ML technologies. Consider the website "Hugging Tweets," for example, in which users can fine-tune a pre-trained GPT-2 language model on their own Twitter stream (or another public user's) in order to see what new tweets the model would suggest for their timeline, based on previous activity.[153] The results are, at least in my own case, sometimes uncannily accurate and often non-sensical.

Whether intended as a creative output or not, using this interface prompts immediate questions about the source data used for pre-training, as well as the assumptions baked into the prediction model. Like Crawford and Paglen's "ImageNet Roulette," actually working with an ML algorithm contributes, even in subtle ways, to goals of algorithmic literacy (see 5.5.1). While many public projects currently rely on data such as Twitter, due to the relative ease of accessing its API, these kind of experiments could be vastly enriched if more historical, humanistic, or social scientific library collections were made available for them. As with all such decisions, libraries must be careful about whose materials are opened and to whom, but I would argue that cultivating more opportunities for reuse, remix, and mash-up of cultural heritage materials will drive public attention and interest. In addition to research-focused ML "data jams," libraries might also sponsor more creative data events and collaborate with ML-engaged artists to produce work derived from digital collections.

[152] Eileen Jakeway, "Newspaper Navigator Surfaces Treasure Trove of Historic Images a Sneak Peek at Upcoming Data Jam! | the Signal," webpage, April 21, 2020, https://blogs.loc.gov/thesignal/2020/04/newspaper-navigator-surfaces-treasure-trove-of-historic-images-get-a-sneak-peek-at-upcoming-data-jam/.

[153] Boris Dayma, "HuggingTweets - Train a Model to Generate Tweets," W&B, 2020, https://tinyurl.com/ybaqmhzo.

5.3. Develop ML + Libraries Infrastructure

The infrastructure required for ML has shifted considerably in the past decade, as I will discuss below. Nevertheless, questions of both hardware and, in particular, staffing dominate librarians' concerns about participating in this new area of research. This is particularly true for smaller institutions, who perhaps lack the capacity for significant investments in order to ascertain whether ML will be useful for their collections.

5.3.1. Create Memoranda of Understanding for ML Project Staffing

At both the ML + Libraries event and in interviews, experts affirmed that successful ML projects clearly establish divisions of labor and expectations for collaborators, while defining specific goals and outcomes for all partners on ML projects. The most challenging projects were those in which the different roles, disciplines, and expertise of the collaborators were not equally considered throughout the life of the project. The most successful projects benefit all collaborators in tangible ways that resonate in their professions or fields, so that no one is asked to donate their time, attention, or intellectual or emotional labor without just recompense. Such recompense need not always be financial. For example, a research scholar might consider a peer-reviewed publication in their field a vital outcome of an ML project. What is most important is that expectations are negotiated and clearly delineated before a project begins, and that all partners regularly check on progress toward these discrete deliverables throughout the life of a project.

Veterans of successful ML projects recommended memoranda of understanding (MOUs) as particularly useful for outlining collective and individual duties, responsibilities, and outcomes for ML projects. A project MOU names the roles each collaborator is expected to assume on a project (e.g. project manager, technical developer), describes the precise responsibilities of each collaborator (e.g. subject area expertise, technical competencies, resources), outlines the practical expectations of timing and reporting for the project members (e.g. mode and frequency of engagement, assessment of contributions, hierarchies of reporting), lists expected outcomes of project work (e.g. publications, datasets, interfaces, visualizations), and defines the conditions of success for each collaborator. MOUs should be as precise as possible, including dates of deliverables, so that all project members understand both what is expected of them and how they stand to benefit professionally from the work. The goal of an MOU is not punitive, but instead to fully recognize the contributions and professionalism of all partners, especially across modes of expertise and discipline, and to mitigate potential imbalances of power across a project team.

5.3.2. Publish and Collect Sample ML Pipelines

As outlined in 4.4, a significant challenge to even beginning ML projects is understanding the full ML workflow, or pipeline, that would move a library from collection through an ML process and interpretation. There is no single answer to this challenge, save that more examples are needed that seek to capture the full pipeline rather than focusing on the final stages. The "higher order" work of those final stages—e.g. computational classification, automatic content tagging—can be

most compelling from a scholarly perspective and demonstrate the potential impact of ML methods, but focus on those stages can obscure the full pipeline required for libraries to get to those stages, leave newcomers to ML methods uncertain how to begin, and convey the notion of ML methods as more objective than they actually are. As sociologist Laura Nelson argued in an interview, whenever results are published from ML methods, researchers should list every subjective decision made to get to those results, including their choice of algorithm(s), and pre-processing steps (e.g. removing stopwords, lower casing words). "The worst thing we could be doing," she continued, "is pretending these are objective presentations of data rather than the results of series of decisions."[154]

I point to several exemplars here that seek to outline the full ML pipeline in a library context, and encourage further publication along these lines. In addition, I urge libraries leading the ML conversation to prioritize documentation and publication of process documents that will help guide their colleagues through ML conceptualization and implementation. A cluster of these exemplars focus on projects using the Library of Congress's Chronicling America newspaper collection. First, the IDA report, cited throughout this document, specifically frames itself as "a demonstration project" and documents each stage of the team's pipeline in careful detail. The code and data for the project are stored in GitHub repositories and linked in the report. Such thorough documentation supports both reproducibility and parallel development for teams that wish to adapt their model to other domains.[155]

Similarly, the Newspaper Navigator team outlines their work building a "visual content recognition model" for Chronicling America newspapers by

> describ[ing] our pipeline that utilizes this deep learning model to extract 7 classes of visual content: headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements, complete with textual content such as captions derived from the METS/ALTO OCR, as well as image embeddings for fast image similarity querying. We report the results of running the pipeline on 16.3 million pages from the Chronicling America corpus and describe the resulting Newspaper Navigator dataset, the largest dataset of extracted visual content from historic newspapers ever produced. The Newspaper Navigator dataset, fine-tuned visual content recognition model, and all source code are placed in the public domain for unrestricted re-use.[156]

This overview of the full Newspaper Navigator pipeline exemplifies the best practices ML projects should adapt moving forward. It outlines the shape of its source data, including crowdsourced annotation data from the LoC's Beyond Words initiative; describes the stages of annotation, correction, and verification required to construct its training sets; details the ML models adapted to recognize this visual content; describes the results of the research; and provides both pipeline and code for testing, reuse, or adaptation by other researchers. As more libraries undertake ML projects, such documentation should be imitated as the gold standard.

These projects using Chronicling America are certainly not the only exemplars for pipeline documentation. I would point to Mühling *et al*'s publication about GDR television broadcasts, described

[154] Cordell, "Machine Learning and Libraries Interview with Laura Nelson."
[155] "Image Analysis for Archival Discovery (Aida)."
[156] Lee et al., "The Newspaper Navigator Dataset," pg. 1.

in 3.2.7, as another detailed and useful pipeline description, and there are many other examples in the bibliography for this report. Essentially, however, these pipeline descriptions remain largely embedded in research publications, scattered through a range of disciplinary journals or repositories. Initially, libraries should cultivate the collection of ML pipeline bibliographies to help staff and external researchers identify models as they plan ML projects.

5.3.3. Develop ML Implementation Toolkits

The workflow bibliographies recommended in 5.3.2, however, only go so far to assist newcomers to ML methods. The specific workflows outlined in the existing literature can be hard to generalize from, particularly for teams new to ML work. To that end, funders and leading ML libraries should support the development of ML Implementation Toolkits for distinct data types (e.g. text, image, audio, video) or in particular domains. These toolkits would include:

1. Model training data, with descriptions of annotation processes undertaken and an intellectual justification of the same.
2. An inventory of existing resources, including open-access, domain-specific collections; ML models, algorithms, and code that could be used or adapted; and prospective pre-training data from earlier ML projects in the domain.
3. A walkthrough of the full technical pipeline from related project, including training, validation, and application of the model.
4. Model code from related experiments.

What I name here as an implementation toolkit generalizes the kinds documentation provided by the demonstration projects I advocate in 5.4.1, such that these goals could dovetail with the sustained development of demonstration projects across media types and domains.

5.4. Support for ML + Library Projects

The ML and libraries conversation will be most substantially advanced by the development of more projects employing ML methods using library collections and publishing their results, including training data, code, and pipelines, for other libraries to emulate. Doing so will require institutions willing to make ML an institutional priority, from the creation of machine-actionable digitized collections; to hiring or training ML-adept staff; to funding responsible and sustainable ML projects, including the algorithmic accountability measures outlined in Section 5.1.4.

5.4.1. Fund Pilot ML Demonstration Projects

While ML experiments have been underway in libraries for decades, the movement toward explainable ML has gained significant traction in the past few years. As a result, there has been a welcome boom of projects such as IDA and Newspaper Navigator, both frequently cited in this report, which present ML-driven analyses employing library collections while clearly documenting their research

processes, technical steps, data, code, and pipelines.[157] IDA describes itself as "a demonstration project," a term that speaks to the most urgent needs of the community right now: models of the full ML and libraries pipeline that can be emulated by other institutions.

In the next two years, leading ML-capable libraries and funders should support 3-5 additional, funded demonstration projects in distinct historical or humanistic domains. In our interview, Kate Zwaard advocated the value of pilot projects to "give everyone a sense of the tangible cost-benefits of setting up and running a[n ML] project."[158] These benefits redound not only to the library sponsoring the project but also, if properly documented, the wider ML and libraries community. Like IDA and Newspaper Navigator, the goal of future demonstration projects should be trifold. First, they should seek to investigate particular datasets using ML methods in ways meaningful to domain scholars, demonstrating the potential value of ML for research in the area. Secondly, they should forefront questions of data bias, harm mitigation, and data justice in their project conception, process descriptions, and analyses. Finally, they should prioritize extensive documentation of their project data, pipeline, and workflows, for the benefit of the wider library community.

If possible, these demonstration projects could be organized and funded under a unified initiative, along the lines of the Digging Into Data Challenges. The goal of such coordination would be to increase communication and idea exchange among the demonstration projects and provide regular opportunities for dialogue through program meetings or conferences. In addition, projects' data, whitepapers, and other outputs could be aggregated and shared together, seeding efforts to build ML and library resource guides, syllabi, and toolkits with salient examples.

5.4.2. Pilot Integrated Interfaces for Communicating ML-Derived Data

As I describe in 4.5, much of the literature in ML and libraries is composed in the "perpetual future tense." Projects are framed in terms of the discoveries they will spark by allowing users to browse or search library data in new ways. The applications outlined in Section 3 of this report exemplify this perpetual future tense. The many activities outlined under the heading of "discoverability" in 3.2— e.g. clustering, classification, metadata exaction—do not become really valuable until users can make use of them to explore. Currently there is a stark divide between most collections and the outputs of the research those collections enable, but this divide could be bridged as ML data is used to directly enrich collections. Projects such as "National Neighbors"[159] and "Neural Neighbors"[160] suggest what ML-integrated interfaces might look like, and the ways they might enable researchers, students, and other users to interact with collections in new ways, but even these interfaces do not directly tie into library catalog systems.

I strongly recommend that libraries strive not for polish in these interfaces, but instead for explicit communication of ML processes and decisions. In our interview, Benjamin Lee identified human-computer interaction (HCI) as one of the most fruitful areas for collaboration between libraries and computer scientists around machine learning. From the perspective of HCI, he argued, a researcher

[157] Lorang et al., "Digital Libraries, Intelligent Data Analytics, and Augmented Description," lee2020a.

[158] Cordell, "Machine Learning and Libraries Interview with Kate Zwaard."

[159] Sarah Reiff Conell et al., "National Neighbors - CMU DH," accessed November 6, 2019, https://dh-web.hss.cmu.edu/nga/essay.

[160] Yale Digital Humanities Lab, "Neural Neighbors."

should not wait for the data to be perfect, but instead present it as a pilot or prototype, learn from users, and refine from there.[161] Where computer scientists bring expertise in ML methods, libraries understand information literacy and can help construct interfaces that communicate ML data to a wide range of prospective users.

ML-integrated interfaces should report, for instance, the confidence scores for relationships, annotations, or other metadata determined through an ML algorithm, and seek to make users more aware of the probabilistic basis of this data. The aim of such reporting is not simply to cast doubt—though skepticism is healthy in this domain—but to make presentations of ML data opportunities for cultivating literacy for ML and probabilistic methods. Helping users understand the confidence rating behind a particular label or category helps contextualize any claims they might make. A sense of ML's limitations could, perhaps counterintuitively, serve to increase overall confidence in ML, because its claims will be understood as contextual and relational rather than totalizing. Through the cultivation of explaining, ML-integrated interfaces, the library will help meet the educational goals I outline in 5.5.1 below, centering its engagements with ML through pedagogy and outreach.

5.5. ML Expertise in Libraries

Beyond serving as sites for responsible ML research projects, libraries have a central role to play in cultivating literacy about ML technologies, and their effects in patrons' lives. While many companies see their duty primarily to their shareholders, libraries have a public mission that gives them a unique opportunity to educate as they experiment. As sites for applied pedagogy, libraries can become transformative advocates in broader cultural and political conversations about ML. Fostering these educational efforts will require infrastructural investments in ML expertise among library staff and creative thinking about how to share expertise among institutions.

5.5.1. Incorporate ML Into Library Information Literacy Pedagogy

In addition to applying ML methods to collections, libraries must incorporate ML into their information literacy pedagogy. Both university and public libraries prioritize literacy and education, for instance training patrons to evaluate the reliability of research sources, whether analog or digital. Given the development of "deep fakes" and language models such as GPT-2, it will become more necessary than ever to educate researchers, students, and the public about the capabilities of AI systems for both positive and malicious purposes. As the developers of GPT-2 acknowledge, such technology could be used, among other things, to:

- Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content

"The public at large," they continue, "will need to become more skeptical of text they find online,

[161] Cordell, "Machine Learning and Libraries Interview with Benjamin Charles Germain Lee."

just as the 'deep fakes' phenomenon calls for more skepticism about images."[162] In such a media environment, libraries must retool many educational programs, especially those aimed at teaching research methods, to help their users become aware of the capabilities for AI/ML materials, better identify such materials, and evaluate the veracity of AI or ML-generated sources. Given the many positive uses for ML outlined in this report, this is not a recommendation for outright refusal, but instead careful assessment of such material that gives users power to use them responsibly and effectively.

In this way, ML education can be part of a broader movement toward "algorithmic literacy." Michael Ridley positions algorithmic literacy alongside his call for explainable AI:

> Research libraries, like all libraries, have been active proponents of enhancing literacy, be it traditional reading and writing or more recently digital literacy in all its various forms. While algorithmic literacy can be seen as a subset of digital literacy or computational thinking, it has unique characteristics and applications that deserve specific attention. Just as information literacy provides users with skills and perspectives to assess resources, algorithmic literacy is an explainability strategy allowing users to navigate and utilize algorithmic tools and services.[163]

In addition to empowering patrons to interact with ML algorithms, algorithmic literacy helps address the cautions overviewed in Section 2 of this report.

5.5.2. Develop Modules for ML Training in MLIS Programs

In order to meet the needs of both ML research projects and ML literacy pedagogy, libraries themselves will require more widespread literacy about ML methods. During the ML + Libraries event, participants proposed the development of ML curricula within MLIS programs to help address this need. Given the wide range of topics MLIS students are asked to learn during a relatively short degree program, it likely isn't feasible to train library students to become ML experts, though there might be room for some MLIS programs to create a specialized ML track.

More immediately, however, MLIS programs could prioritize training in "computational thinking" that will allow library students to understand the broad scope of ML methods and their applications, learn where they might acquire specialized knowledge needed for particular projects, and, most importantly, join teams of experts as effective collaborators. Every librarian need not be capable of taking on the role of central technologist for ML projects, but a broad infusion of ML literacy would prove useful across multiple kinds of engagement with ML in library work. Given the rapid evolution of ML technologies, this "soft skills" approach may prove more durable than mechanistic instruction in particular ML platforms or tools.

In practical terms, then, there may be some opportunities for MLIS programs to implement entire classes in ML methods. However, most programs might first develop ML modules that can be integrated across existing curricula. Funders should consider a challenge grant that would bring

[162] Alec Radford et al., "Better Language Models and Their Implications," OpenAI, accessed November 13, 2019, https://openai.com/blog/better-language-models/.

[163] Ridley, "Explainable Artificial Intelligence (RLI 299, 2019)," pg. 39.

together faculty from several MLIS programs to develop a pilot module along these lines for deployment at several institutions over the next 2-3 years. This module should demonstrate ML methods grounded in library-specific use cases. This module could also begin to serve the professional development recommendations outlined in the next section (5.5.3) and provide a model for the development of ML classes or course sequences. Over the next 5 years, funders and institutions should support the development of a set of modules training library staff in responsible ML methods, grounded in library use cases, for deployment in MLIS programs and professional development initiatives.

5.5.3. Cultivate Opportunities for Professional Development in ML

As stated above, library staffing needs in ML are unlikely to be fully met by training of new librarians in MLIS programs: current needs outpace the training possible during a two-year degree program and the existing expertise gap is wider than the influx of new librarians could bridge. As outlined in 4.2, outside ML experts generally command significant salaries that make it challenging for libraries to hire full-time ML expertise. As Padilla argues in RO, "Arguments that libraries can secure the talent they need by virtue of the distinctiveness of their mission are flattened by the reality of the rising cost of living throughout the US. Increasing the number of staff with these capabilities across an organization moves the recruitment and retention of staff with highly sought-after technical skills from an edge case to a core concern."[164] This expertise gap can be addressed somewhat through collaboration, particularly in academic contexts where library collections might advance the research goals of colleagues in computer or data science. But limited-term collaborations risk turning into boutique projects that disappear when the core collaborators turn to new projects. Some form of staff continuation is necessary to structure ML projects through the steps of conceptualization, execution, and integration suggested throughout this report.

In that vein, I would amplify RO's suggestion that libraries commit to "developing internal talent through mentoring programs, education, experiential opportunities, and clear paths to making use of what they learn without the threat of it stacking onto existing job responsibilities." RO highlights the need for evidence-based training and points to a need for training that is "grounded in library use cases."[165] That latter point is key. Many freely-available computational training resources (e.g. Khan Academy, Coursera) are grounded in particular academic or corporate contexts and center on examples that can be difficult to extrapolate into library, humanistic, or cultural heritage contexts. The act of translation required to move from the lessons available through such platforms to the specific applications in library contexts can stymy even the best-intended professional.

More tailored opportunities will be needed to foster ML expertise among librarians and related staff. Workshops and working groups may help generate ideas for ML work, but such limited engagements are not solutions for substantial expertise development. Broader institutional commitments to training will be required to sustain projects, as well as to undertake the regular audits recommended for responsible ML work, as recommended above. If funded, the training modules

[164] Padilla, "Responsible Operations," pg. 18.
[165] Padilla, pg. 18-19.

suggested for MLIS programs in 5.5.2 could help support development opportunities for current library staff, particularly if integrated into ongoing, funded professional development.

5.5.4. Pool ML Expertise

Considering the broader library and cultural heritage landscape, it will likely never be feasible for all institutions to maintain full-time ML expertise locally. Participants at the ML + Libraries event discussed a range of strategies for collective hiring and pooling expertise in order to address these needs. For some strong library consortia, shared hiring might begin to address these needs, but project managers might also consider broader communication around funding proposals, such as grants. Institutions holding datasets in the same domain might apply for joint funding, for instance, to support needed ML expertise.

More ambitiously, Dan Cohen suggested that a small non-profit organization might be created to contract with libraries in need of ML support. Libraries could pay a service fee in exchange for design and ML expertise, paired with a commitment of in-kind labor at the contracting library.[166] Such an operation would bridge a significant gap in the ML conversation, where big companies such as Google or Microsoft are unlikely to be interested in the relatively small data and ML needs of libraries, museums, or related institutions, while the technical needs of those institutions still exceed available funding and staff.

5.5.5. Develop Guidelines for Vendor Solutions

This report has deliberately avoided much discussion of vendor-supplied ML tools, primary because I do not, in general, believe they meet the standards of openness, explainability, and adaptability that best practices encourage. While vendor-built tools can be attractive, particularly given the challenges of ML expertise outlined in 4.2, these systems can ultimately restrict as much as they enable. As Laura Nelson argued, "creating"—or, I would add, subscribing to—"GUI [graphic user interface] tool after GUI tool is a waste of time and money. Too often you get stuck and realize it doesn't do what you want it to do." For Nelson, the better path for libraries is to make more collections available to researchers who can work with their data, and to "partner…with open source communities, training people to use those tools."[167] This recommendation amplifies this report's focus on the library as a site for collaboration and ML pedagogy.

The Library of Congress's Abigail Potter expressed similar concerns to Nelson, while acknowledging that some libraries will use vendor solutions "because there isn't a ton of experience inside libraries and because the commercial marketplace for ML and AI services seems pretty hot right now." However, Potter amplified Nelson's concerns that "there is a mismatch in what is being offered, i.e., full solutions or black box tools, and the needs in cultural heritage for transparency, assessment, auditing and perhaps reprocessing of data." Potter recommended that the library community develop "guidelines about how to evaluate or perform quality assurance on vendor-offered

[166] Cordell, "Machine Learning and Libraries Interview with Dan Cohen."
[167] Cordell, "Machine Learning and Libraries Interview with Laura Nelson."

solutions, like the FADGI (Federal Agency Digitization Guidelines Initiative) guidelines."[168] This is another area in which leading libraries such as the LoC could take the lead, convening a working group to begin drafting these assessment guidelines, likely based on the statement of values and algorithmic assessment toolkits recommended earlier in this report.

6. 25 Questions for Structuring an ML Project

In 5.1.4, I quote Rachel Thomas about the value of checklists for "help[ing] those working in tech make fewer ethical mistakes."[169] In that same piece, Thomas provides an example of such a checklist, which I quote in full below as a useful model:

- Have we listed how this technology can be attacked or abused?
- Have we tested our training data to ensure that it is fair and representative?
- Have we studied and understood possible sources of bias in our data?
- Does our team reflect diversity of opinions, backgrounds, and kinds of thought?
- What kinds of user consent do we need to collect or use the data?
- Do we have a mechanism for gathering consent from users?
- Have we explained clearly what users are consenting to?
- Do we have a mechanism for redress if people are harmed by the results?
- Can we shut down this software in production if it is behaving badly?
- Have we tested for fairness with respect to different user groups?
- Have we tested for disparate error rates among different user groups?
- Do we test and monitor for model drift to ensure our software remains fair over time?[170]

Several of Thomas' questions could transfer directly to checklists for responsible ML development in libraries. I will adapt some of these in the document below.

However, the specifics of the cultural heritage domain require some amendments and additions to this checklist. In the following section of this report, I provide a series of questions—drawing on the literature, challenges, and recommendations outlined in previous sections—to help library teams structure responsible and responsive ML projects. This list does not pretend to be comprehensive, and thus its first question is simply:

1. What new, amended, or special considerations pertain to our ML project that are not covered in the Library of Congress report's model checklist?

I do not provide a full rationale for each of the questions below, but instead cite pertinent sections of this report where you can find the explanations, examples, and citations that underlie each item.

[168] Potter, "Report Draft Comments."
[169] Thomas, "16 Things You Can Do to Make Tech More Ethical, Part 1."
[170] Thomas.

6.1. Values

2. Has our library drafted a statement of values (SoV) for ML? Does it reflect our current values, as needed for this project, or does it need updating? (5.1.3)

3. If we do not have an SoV, are there models from peer institutions we can adopt or adapt? (5.1.2)

4. Have we developed algorithmic impact assessments and a plan for explainable ML? (5.1.4, 5.1.5)

5. What will be the environmental consequences of this project, and can we offset those impacts in meaningful ways? (2.3)

6. What is our plan to revisit and (if necessary) adapt our SoV as the project develops? (5.1.6)

6.2. Staff and Expertise

7. Do we have the staff expertise necessary for this ML project? Can existing project staff manage data acquisition, cleaning, and management; domain-specific training data annotation; ML development and deployment; interface and visualization development; domain-specific evaluation of results? (4.2)

8. If we do not have the necessary expertise, can we train from within to meet this project's needs, thereby (usefully) enriching a team member's portfolio while contributing to our institutional goals? (5.5.3)

9. If training from within is not possible or desirable, can we find opportunities to collaborate with other institutions to meet the needs of this project? Does this project give us an opportunity to pool expertise and work collectively? (5.5.4)

10. Does our team reflect diversity of opinions, backgrounds, and kinds of thought?[171] If not, how might we proactively correct this lack? (2.1)

6.3. Collaboration and Team Expectations

11. Have we drafted a memorandum of understanding (MOU) outlining the duties of all collaborators and anticipated outcomes, both collective and individual, for the project work? (5.3.1)

12. Can we ensure that the skills and expertise of all project participants are valued, and that we have set appropriate outcomes so that all collaborators are fairly recognized and compensated for their contributions? (3.1, 5.3.1)

[171] Thomas.

6.4. Hardware and Software

13. Does the library already have adequate hardware to train ML models and apply them across the desired collections? (4.3)

14. Do we have the correct software environments set up to run ML algorithms? What software do we need set up, and do we have the necessary expertise and hardware to do so? (4.3)

15. If the answer to either of the above is "no," can we collaborate with another institution to establish needed capacities?

16. If the answer to the above is "no," do we have a budget to establish necessary hardware and/or software, or to rent the necessary equipment/space?

6.5. ML Data

17. Is there machine-actionable data in the domain we seek to investigate? If not, do we have a plan for digitization of collections? (4.1.1, 5.2.1)

18. Does training data exist in the domain we seek to model? If not, do we have the domain expertise to annotate training data? (4.1.2, 5.2.2)

19. If training data is not available for the domain we seek to model, is there similar training data we might use to pre-train a ML model that can then be fine-tuned with a smaller set of domain-specific data? (4.1.4, 5.2.2)

20. Have we audited our training data to ensure it is representative and to mitigate inevitable biases of availability, selection, and expertise? (2.1, 4.1.3)

21. Are there communities of experts or citizen researchers we might meaningfully involve in annotating training data, evaluating results, or piloting interfaces and visualizations? (5.2.3)

6.8. Outcomes

22. Do we have a clear pipeline established to move from project conception to sustainable implementation of ML results? (5.3.2)

23. Can we imagine interfaces or visualizations that model explainable ML and communicate the results of our work to patrons? (5.1.5, 5.5.1)

24. Can we publish all training data, code, and ML-annotated data for the benefit of other ML researchers? If not, what infrastructure would be required to publish these elements? (5.2.2)

25. Can our ML-annotated data be incorporated into our systems for collection exploration and discovery? In other words, will this experiment enhance our collections in the ways we hope? (5.4.2)

7. Appendices

Appendix A: Bibliography

This report bibliography is also available as a Zotero Group library at https://www.zotero.org/groups/2497390/machine_learning__libraries/library.

Abebe, Rediet, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. "Roles for Computing in Social Change." *ArXiv:1912.04883 [Cs]*, January 28, 2020. https://doi.org/10.1145/3351095.3372871.

Adams, Nick. "Library of Congress: Archival Search and Discovery, Machine Learning, and Stoicism [Recap]." TagWorks, September 25, 2019. https://tag.works/blog/2019/9/29/library-of-congress-recap-archival-search-and-discovery-machine-learning-and-stoicism.

———. "NLP, AI, and Social Science Are About to Get A Lot Better." Medium, August 7, 2019. https://medium.com/@nick_65591/ai-and-social-science-are-about-to-get-a-lot-better-6e3c07a44502.

AI Now. "Algorithmic Accountability Policy Toolkit," October 2018.

Alammar, Jay. "The Illustrated GPT-2 (Visualizing Transformer Language Models)," August 12, 2019. http://jalammar.github.io/illustrated-gpt2/.

———. "The Illustrated Transformer," June 27, 2018. http://jalammar.github.io/illustrated-transformer/.

———. "The Illustrated Word2vec," March 27, 2019. http://jalammar.github.io/illustrated-word2vec/.

———. "Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)," May 9, 2018. https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/.

Algorithmic Justice League. "Safe Face Pledge." Center on Privacy & Technology at Georgetown Law, 2018. https://www.safefacepledge.org/.

Allen Institute for AI. "AllenNLP - Demo." Accessed February 20, 2020. https://demo.allennlp.org/next-token-lm?text=Joel%20is.

Alpaydin, Ethem. *Machine Learning*. The MIT Press Essential Knowledge Series. The MIT Press, 2016. https://mitpress.mit.edu/books/machine-learning.

———. "Machine Learning." The MIT Press. Accessed July 24, 2019. https://mitpress.mit.edu/books/machine-learning.

Alpert-Abrams, Hannah. "Machine Reading the Primeros Libros" 10, no. 4 (2016). http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html.

Alvarez Melis, David, and Tommi Jaakkola. "Towards Robust Interpretability with Self-Explaining Neural Networks." In *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H.

Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 7775–7784. Curran Associates, Inc., 2018. http://papers.nips.cc/paper/8003-towards-robust-interpretability-with-self-explaining-neural-networks.pdf.

Amaro, Ramon. "As If." *Becoming Digital* (blog). Accessed January 17, 2020. https://www.e-flux.com/architecture/becoming-digital/248073/as-if/.

American Library Association. "Power That Is Moral: Cataloging and Ethics." Text. Association for Library Collections & Technical Services (ALCTS), July 25, 2017. http://www.ala.org/alcts/confevents/upcoming/e-forum/090517.

Amnesty International, and Access Now. "The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems," May 16, 2018. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

Amodei, Dario, Danny Hernandez, Girish Sastry, Jack Clark, Greg Brockman, and Ilya Sutskever. "AI and Compute." OpenAI, November 7, 2019. https://openai.com/blog/ai-and-compute/.

Angwin, Julia, and Jeff Larson. "Machine Bias." Text/html. ProPublica, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ardanuy, Mariona Coll, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. "Living Machines: A Study of Atypical Animacy." *ArXiv:2005.11140 [Cs]*, May 22, 2020. http://arxiv.org/abs/2005.11140.

Arnold, Taylor. "Replication Data for: 'A Visual Style in Two Network Sitcoms' by Taylor Arnold, Lauren Tilton, and Annie Berke." Harvard Dataverse, 2019. https://doi.org/10.7910/dvn/s84tsx.

Arnold, Taylor, Lauren Tilton, and Annie Berke. "Visual Style in Two Network Era Sitcoms." *Journal of Cultural Analytics*, 2019. https://doi.org/10.22148/16.043.

"Artificial Intelligence and Machine Learning in Libraries | ALA Store." Library Technology Reports. American Library Association, 2019. https://www.alastore.ala.org/content/artificial-intelligence-and-machine-learning-libraries.

"Artificial Intelligence Has a 'Sea of Dudes' Problem." *Bloomberg Professional Services*, June 27, 2016, sec. Import - Enterprise (3-16-2017). https://www.bloomberg.com/professional/blog/artificial-intelligence-sea-dudes-problem/.

Arya, Vijay, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, et al. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." *ArXiv:1909.03012 [Cs, Stat]*, September 14, 2019. http://arxiv.org/abs/1909.03012.

Ayre, Lori, and Jim Craner. "Algorithms: Avoiding the Implementation of Institutional Biases." *Public Library Quarterly* 37, no. 3 (July 3, 2018): 341–47. https://doi.org/10.1080/01616846.2018.1512811.

Basaran, Dogac, Slim Essid, and Geoffroy Peeters. "MAIN MELODY EXTRACTION WITH SOURCE-FILTER NMF AND CRNN." In *19th International Society for Music Information Retreival*.

Paris, France, 2018. https://hal.archives-ouvertes.fr/hal-02019103.

Beals, M. H., and Emily Bell. "The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges." Loughborough, 2020. https://www.digitisednewspapers.net/.

Bell, Eamonn, and Laurent Pugin. "Heuristic and Supervised Approaches to Handwritten Annotation Extraction for Musical Score Images." *International Journal on Digital Libraries* 20, no. 1 (March 1, 2019): 49–59. https://doi.org/10.1007/s00799-018-0249-7.

Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, 2019.

Blanke, Tobias. "Predicting the Past." *Digital Humanities Quarterly* 012, no. 2 (July 3, 2018). http://digitalhumanities.org:8081/dhq/vol/12/2/000377/000377.html.

Blanke, Tobias, Michael Bryant, and Mark Hedges. "Understanding Memories of the Holocaust—A New Approach to Neural Networks in the Digital Humanities." *Digital Scholarship in the Humanities* 0, no. 0 (2019). https://doi.org/10.1093/llc/fqy082.

Blevins, Cameron, and Lincoln Mullen. "Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction." *Digital Humanities Quarterly* 009, no. 3 (December 23, 2015). http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html.

Boer, Victor de, Roeland J. F. Ordelman, and Josefien Schuurman. "Evaluating Unsupervised Thesaurus-Based Labeling of Audiovisual Content in an Archive Production Environment." *International Journal on Digital Libraries* 17, no. 3 (September 1, 2016): 189–201. https://doi.org/10.1007/s00799-016-0182-6.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," 9. Barcelona, Spain, 2016.

Bourg, Chris. "What Happens to Libraries and Librarians When Machines Can Read All the Books?" *Feral Librarian* (blog), March 17, 2017. https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/.

Brygfjeld, Svein Arne, Freddy Wetjen, and André Walsøe. "Machine Learning for Production of Dewey Decimal," July 19, 2018, 9.

Budds, Diana, Diana Budds, and Diana Budds. "Biased AI Is A Threat To Civil Liberties. The ACLU Has A Plan To Fix It." Fast Company, July 25, 2017. https://www.fastcompany.com/90134278/biased-ai-is-a-threat-to-civil-liberty-the-aclu-has-a-plan-to-fix-it.

Buolamwini, Joy. "The Coded Gaze." AJL -ALGORITHMIC JUSTICE LEAGUE, November 6, 2016. https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148.

Byrum, Joseph. "Build a Diverse Team to Solve the AI Riddle." MIT Sloan Management Review, May 18, 2020. https://sloanreview.mit.edu/article/build-a-diverse-team-to-solve-the-ai-riddle/.

Calderon, Ania, Dan Taber, Hong Qu, and Jeff Wen. "AI Blindspot: A Discovery Process for Preventing, Detecting, and Mitigating Bias in AI Systems," 2019. https://aiblindspot.media.mit.edu/.

Caplan, Robyn, Lauren Hanson, and Jeanna Matthews. "Algorithmic Accountability: A Primer." Data & Society, April 18, 2018. https://datasociety.net/output/algorithmic-accountability-a-primer/.

Chen, Hsinchun, Terence R. Smith, Mary L. Larsgaard, Linda L. Hill, and Marshall Ramsey. "A Geographic Knowledge Representation System for Multimedia Geospatial Retrieval and Analysis." *International Journal on Digital Libraries* 1, no. 2 (September 1, 1997): 132–52. https://doi.org/10.1007/s007990050010.

City of Helsinki. "Helsinki City Library Will Be Introducing an AI-Based Intelligent Material Management System | City of Helsinki." City of Helsinki, May 14, 2019. https://www.hel.fi/uutiset/en/kulttuurin-ja-vapaa-ajan-toimiala/helsinki-city-library-wil-be-introducing-an-ai-based-intelligent-material-management-system.

"Civil War Photo Sleuth." Accessed May 4, 2020. https://www.civilwarphotosleuth.com/.

Clement, Tanya, David Tcheng, Loretta Auvil, Boris Capitanu, and Megan Monroe. "Sounding for Meaning: Using Theories of Knowledge Representation to Analyze Aural Patterns in Texts." *Digital Humanities Quarterly* 007, no. 1 (July 1, 2013). http://digitalhumanities.org:8081/dhq/vol/7/1/000146/000146.html.

Cohen, Dan. Machine Learning and Libraries Interview with Dan Cohen. Interview by Ryan Cordell, February 7, 2020.

Colavizza, Giovanni. "Are We Breaking the Social Contract? « CA: Journal of Cultural Analytics." *Journal of Cultural Analytics*, September 17, 2019. https://culturalanalytics.org/article/11828.

Coleman, Catherine Nicole. "Artificial Intelligence and the Library of the Future, Revisited." Stanford Libraries, November 3, 2017. https://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited.

Cordell, Ryan. " 'Q i-Jtb the Raven': Taking Dirty OCR Seriously." *Book History* 20 (2017): 188–225.

Cordell, Ryan, and David Smith. "Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines." Digital History Project. The Viral Texts Project, 2020. https://viraltexts.org.

Costabile, M. F., F. Esposito, G. Semeraro, and N. Fanizzi. "An Adaptive Visual Environment for Digital Libraries." *International Journal on Digital Libraries* 2, no. 2 (September 1, 1999): 124–43. https://doi.org/10.1007/s007990050042.

Crawford, Kate, and Vladan Joler. "Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor." AI Now Institute and Share Lab, September 7, 2018. http://www.anatomyof.ai.

Crawford, Kate, and Trevor Paglen. "Excavating AI: The Politics of Training Sets for Machine Learning." -, September 19, 2019. https://www.excavating.ai.

Critical Platform Studies Group. "Algorithmic Equity Toolkit." ACLU of Washington, April 2020. https://www.aclu-wa.org/AEKit.

Davis, Hannah. "TransProse." TransProse, 2013. http://www.musicfromtext.com/.

Dayma, Boris. "HuggingTweets - Train a Model to Generate Tweets." W&B, 2020. https://tinyurl.com/ybaqmhzo.

Deines, Nathaniel, Melissa Gill, Matthew Lincoln, and Marissa Clifford. "Six Lessons Learned from Our First Crowdsourcing Project in the Digital Humanities." *The Getty Iris* (blog), February 7, 2018. http://blogs.getty.edu/iris/six-lessons-learned-from-our-first-crowdsourcing-project-in-the-digital-humanities/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*, May 24, 2019. http://arxiv.org/abs/1810.04805.

Dick, Stephanie. "Artificial Intelligence," June 14, 2019. https://doi.org/10.1162/99608f92.92fe150c.

"Dig That Lick," 2019. http://dig-that-lick.eecs.qmul.ac.uk/Dig%20That%20Lick_About.html.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Cambridge, Massachusetts: The MIT Press, 2020.

Doshi-Velez, Finale, and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." *ArXiv:1702.08608 [Cs, Stat]*, March 2, 2017. http://arxiv.org/abs/1702.08608.

Earl, Chandra, Alexander White, Michael Trizna, Paul Frandsen, Akito Kawahara, Sean Brady, and Rebecca Dikow. "Discovering Patterns of Biodiversity in Insects Using Deep Machine Learning." *Biodiversity Information Science and Standards* 3 (July 2, 2019): e37525. https://doi.org/10.3897/biss.3.37525.

Elliott, Devon, and William J. Turkel. "Faster than the Eye:: Using Computer Vision to Explore Sources in the History of Stage Magic." In *Seeing the Past with Computers*, edited by KEVIN KEE and TIMOTHY COMPEAU, 83–94. Experiments with Augmented Reality and Computer Vision for History. University of Michigan Press, 2019. https://www.jstor.org/stable/j.ctvnjbdr0.8.

Emerging Technology from the arXiv. "AI Tackles the Vatican's Secrets." MIT Technology Review, March 15, 2018. https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/s/610530/ai-tackles-the-vaticans-secrets/amp.

Esposito, Floriana, Donato Malerba, Giovanni Semeraro, Nicola Fanizzi, and Stefano Ferilli. "Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service." *International Journal on Digital Libraries* 2, no. 1 (October 1, 1998): 3–19. https://doi.org/10.1007/s007990050033.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.

Extance, Andy. "How AI Technology Can Tame the Scientific Literature." *Nature Toolbox* (blog), September 10, 2018. https://www.nature.com/articles/d41586-018-06617-5.

Fagan, Benjamin. "Chronicling White America." *American Periodicals: A Journal of History & Criticism* 26, no. 1 (2016): 10–13.

Ferris, Anna M. "The Ethics and Integrity of Cataloging." *Journal of Library Administration* 47, no. 3–4 (July 1, 2008): 173–90. https://doi.org/10.1080/01930820802186514.

Ferriter, Meghan. "Introducing Beyond Words." Webpage. The Signal, September 28, 2017. //blogs.loc.gov/thesignal/2017/09/introducing-beyond-words/.

———. "Report Draft Comments," May 15, 2020.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." *SSRN Electronic Journal*, 2020. https://doi.org/10.2139/ssrn.3518482.

Floridi, Luciano, and Josh Cowls. "A Unified Framework of Five Principles for AI in Society," June 14, 2019. https://doi.org/10.1162/99608f92.8cd550d1.

Gallagher, Liz. "Finding Image Pathways." Medium, September 5, 2019. https://stacks.wellcomecollection.org/finding-image-pathways-12d31ae347f9.

Garcia-Febo, Loida. "Exploring AI: How Libraries Are Starting to Apply Artificial Intelligence in Their Work." *American Libraries* 50, no. 3/4 (April 3, 2019): 4–4.

Gavin, Michael. "Is There a Text in My Data? (Part 1): On Counting Words « CA: Journal of Cultural Analytics." *Journal of Cultural Analytics*, September 17, 2019. https://culturalanalytics.org/article/11830.

Glaser, April. "Artificial Intelligence Uses Too Much Energy to Train. Researchers Want to Fix That." Slate Magazine, September 20, 2019. https://slate.com/technology/2019/09/artificial-intelligence-climate-change-carbon-emissions-roy-schwartz.html.

Google. "Artificial Intelligence at Google: Our Principles." Google AI. Accessed September 20, 2019. https://ai.google/principles/.

Google TensorFlow. *Writing the Playbook for Fair & Ethical Artificial Intelligence & Machine Learning (Google I/O'19)*, 2019. https://www.youtube.com/watch?v=5pMQGT3O4CI.

Gotterbarn, Don, Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Vazansky, Marty J Wolf, et al. "ACM Code of Ethics and Professional Conduct," 2018, 28.

Griffey, Jason. "AI and Machine Learning: The Challenges of Artificial Intelligence in Libraries." *American Libraries* 50, no. 3/4 (April 3, 2019): 47–47.

Groth, Olaf J., Mark J. Nitzberg, and Stuart J. Russell. "AI Algorithms Need FDA-Style Drug Trials." *Wired*. Accessed September 16, 2019. https://www.wired.com/story/ai-algorithms-need-drug-trials/.

Guiliano, Jennifer, and Carolyn Heitman. "Difficult Heritage and the Complexities of Indigenous Data." *Journal of Cultural Analytics*, 2019. https://doi.org/10.22148/16.044.

Guldi, Jo. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora." *Journal of Cultural Analytics*, 2018. https://doi.org/10.22148/16.030.

Hao, Karen. "In 2020, Let's Stop AI Ethics-Washing and Actually Do Something." MIT Technology Review. Accessed January 5, 2020. https://www.technologyreview.com/s/614992/ai-ethics-

washing-time-to-act/.

———. "Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes." MIT Technology Review. Accessed October 15, 2019. https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/.

Henry, Geneva. "Research Librarians as Guides and Navigators for AI Policies at Universities (RLI 299, 2019)." *Research Library Issues*, no. 299 (2019). https://publications.arl.org/18nm1dh/.

Hill, Mark J., and Simon Hengchen. "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study." *Digital Scholarship in the Humanities* 34, no. 4 (December 1, 2019): 825–43. https://doi.org/10.1093/llc/fqz024.

Hind, Michael, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. "TED: Teaching AI to Explain Its Decisions." *ArXiv:1811.04896 [Cs]*, November 12, 2018. http://arxiv.org/abs/1811.04896.

Horton, Russell, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie." *Digital Humanities Quarterly* 003, no. 2 (June 18, 2009).

"How It Works | Google Cloud." Accessed September 17, 2019. https://cloud.withgoogle.com/nytimes/how-it-works.

Hu, Margaret. "Algorithmic Jim Crow." *Fordham Law Review* 86, no. 2 (2017): 65.

IBM. "AI Explainability 360 Open Source Toolkit." GitHub, 2019. http://aix360.mybluemix.net/.

"Image Analysis for Archival Discovery (Aida)," 2020. http://projectaida.org/.

"In Codice Ratio." Accessed February 17, 2020. http://www.inf.uniroma3.it/db/icr/.

INDIGENOUS AI. "INDIGENOUS AI," 2019. http://www.indigenous-ai.net/.

Institute, AI Now. "AI in 2019: A Year in Review." Medium, October 9, 2019. https://medium.com/@AINowInstitute/ai-in-2019-a-year-in-review-c1eba5107127.

Cooper Hewitt Smithsonian Design Museum. "Interaction Lab | Cooper Hewitt, Smithsonian Design Museum," September 16, 2019. https://www.cooperhewitt.org/interaction-lab/.

Jacknis, Norman. "The AI-Enhanced Library." Medium, June 21, 2017. https://medium.com/@NormanJacknis/the-ai-enhanced-library-a34d96fffdfe.

Jakeway, Eileen. "Newspaper Navigator Surfaces Treasure Trove of Historic Images – Get a Sneak Peek at Upcoming Data Jam! | The Signal." Webpage, April 21, 2020. https://blogs.loc.gov/thesignal/2020/04/newspaper-navigator-surfaces-treasure-trove-of-historic-images-get-a-sneak-peek-at-upcoming-data-jam/.

Jakeway, Eileen, Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, Abigail Potter, and Kate Zwaard. "Machine Learning + Libraries Summit Event Summary." LC Labs Digital Strategy Directorate, February 13, 2020. https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

Jo, Eun Seo, and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." *ArXiv:1912.10389 [Cs]*, December 22, 2019. https://doi.org/10.1145/3351095.3372829.

Jockers, M. L., and D. M. Witten. "A Comparative Study of Machine Learning Methods for Authorship Attribution." *Literary and Linguistic Computing* 25, no. 2 (June 1, 2010): 215–23. https://doi.org/10.1093/llc/fqq001.

Johnson, Sylvester A. "Technology Innovation and AI Ethics (RLI 299, 2019)." *Research Library Issues*, no. 299 (2019): 14–27. https://doi.org/10.29242/rli.299.2.

Kangas, Pirjo. "About AI for librarians." AI for Librarians. Accessed November 8, 2019. https://www.aiforlibrarians.com/.

———. "Don't Fear AI." *American Libraries* 50, no. 11/12 (December 11, 2019): 7–7.

Katell, Michael, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Binz, Daniella Raz, and P. M. Krafft. "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 45–55. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372874.

Kean, Sam. "Artificial Intelligence Is Cracking Open the Vatican's Secret Archives." The Atlantic, April 30, 2018. https://www.theatlantic.com/technology/archive/2018/04/vatican-secret-archives-artificial-intelligence/559205/.

Kennedy, Mary Lee. "What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries? (RLI 299, 2019)." *Research Library Issues*, no. 299 (2019): 3–13. https://doi.org/10.29242/rli.299.1.

Kennedy, Mary Lee, and Association of Research Libraries. "What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries? (RLI 299, 2019)," September 19, 2019. https://publications.arl.org/rli299/3.

Keralis, Spencer D. C. "Milking the Deficit Internship." In *Disrupting the Digital Humanities*, edited by Dorothy Kim and Jesse Stommel, 2016. http://www.disruptingdh.com/milking-the-deficit-internship/.

Kesserwan, Karina. "How Can Indigenous Knowledge Shape Our View of AI?" Policy Options, February 16, 2018. https://policyoptions.irpp.org/magazines/february-2018/how-can-indigenous-knowledge-shape-our-view-of-ai/.

Keyes, Os. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (November 1, 2018): 1–22. https://doi.org/10.1145/3274357.

———. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (November 1, 2018): 1–22. https://doi.org/10.1145/3274357.

King, Adam. "Talk to Transformer," November 5, 2019. https://talktotransformer.com/?utm_source=dancohen&utm_medium=email.

King, Rachael Scarborough. *Writing to the World: Letters and the Origins of Modern Print Genres*. Baltimore: Johns Hopkins University Press, 2018.

Klampfl, Stefan, Michael Granitzer, Kris Jack, and Roman Kern. "Unsupervised Document Structure Analysis of Digital Scientific Articles." *International Journal on Digital Libraries* 14, no. 3 (August 1, 2014): 83–99. https://doi.org/10.1007/s00799-014-0115-1.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. "OpenNMT: Open-Source Toolkit for Neural Machine Translation." In *Proceedings of ACL 2017, System Demonstrations*, 67–72. Vancouver, Canada: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/P17-4012.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. "Counterfactual Fairness." *ArXiv:1703.06856 [Cs, Stat]*, March 8, 2018. http://arxiv.org/abs/1703.06856.

Laban, Philippe, and Marti Hearst. "NewsLens: Building and Visualizing Long-Ranging News Stories." In *Proceedings of the Events and Stories in the News Workshop*, 1–9. Vancouver, Canada: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/W17-2701.

Lankes, R. David. "Decoding AI and Libraries," July 3, 2019. https://davidlankes.org/decoding-ai-and-libraries/.

Lavin, Matthew J. "Gender Dynamics and Critical Reception: A Study of Early 20th-Century Book Reviews from The New York Times." *Journal of Cultural Analytics*, January 30, 2020, 11831. https://doi.org/10.22148/001c.11831.

Leavy, Susan. "Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning." In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16. GE '18. Gothenburg, Sweden: Association for Computing Machinery, 2018. https://doi.org/10.1145/3195570.3195580.

———. "Uncovering Gender Bias in Newspaper Coverage of Irish Politicians Using Machine Learning." *Digital Scholarship in the Humanities* 34, no. 1 (April 1, 2019): 48–63. https://doi.org/10.1093/llc/fqy005.

Lee, Benjamin Charles Germain. Machine Learning and Libraries Interview with Benjamin Charles Germain Lee. Interview by Ryan Cordell, February 6, 2020.

———. "Machine Learning, Template Matching, and the International Tracing Service Digital Archive: Automating the Retrieval of Death Certificate Reference Cards from 40 Million Document Scans." *Digital Scholarship in the Humanities* 34, no. 3 (September 1, 2019): 513–35. https://doi.org/10.1093/llc/fqy063.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. "The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America." *ArXiv:2005.01583 [Cs]*, May 4, 2020. http://arxiv.org/abs/2005.01583.

Lee, Maurice S. "Falsifiability, Confirmation Bias, and Textual Promiscuity." *J19: The Journal of Nineteenth-Century Americanists* 2, no. 1 (April 3, 2014): 162–71. https://doi.org/10.1353/jnc.2014.0014.

Lewis, Jason Edward, Noelani Arista, Archer Pechawis, and Suzanne Kite. "Making Kin with the Machines." *Journal of Design and Science*, July 16, 2018. https://doi.org/10.21428/bfafd97b.

Library of Congress. "Beyond Words," 2020. http://beyondwords.labs.loc.gov/#/.

———. "Digital Strategy," April 26, 2019. https://www.loc.gov/digital-strategy.

———. *Speech-To-Text Viewer*. Python. 2019. Reprint, Library of Congress, 2020. https://github.com/LibraryOfCongress/speech-to-text-viewer.

library-policy. "The Robots Are Coming? Libraries and Artificial Intelligence « Library Policy and Advocacy Blog." *Library Policy and Advocacy Blog* (blog). Accessed July 29, 2019. http://blogs.ifla.org/lpa/2018/07/24/the-robots-are-coming-libraries-and-artificial-intelligence/.

Lincoln, Matthew D., Golan Levin, Sarah Reiff Conell, and Lingdong Huang. "National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections," November 2019, 20.

Linder, Courtney. "The Librarians of the Future Will Be AI Archivists." Popular Mechanics, May 13, 2020. https://www.popularmechanics.com/technology/a32436235/library-of-congress-machine-learning-newspaper-images/.

Litsey, Ryan, and Weston Mauldin. "Knowing What the Patron Wants: Using Predictive Analytics to Transform Library Decision Making." *The Journal of Academic Librarianship* 44, no. 1 (January 1, 2018): 140–44. https://doi.org/10.1016/j.acalib.2017.09.004.

Lorang, Elizabeth. Machine Learning and Libraries Interview with Elizabeth Lorang. Interview by Ryan Cordell, February 21, 2020.

Lorang, Prepared Elizabeth, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project," January 10, 2020, 43.

Loukides, Mike, Hilary Mason, and D. J. Patil. *Ethics and Data Science*. 1 edition. O'Reilly Media, 2018.

Lynch, Clifford A. "Machine Learning, Archives and Special Collections: A High Level View," October 2, 2019. https://blog-ica.org/2019/10/02/machine-learning-archives-and-special-collections-a-high-level-view/.

Mac, Ryan, Caroline Haskins, and Logan McDonald. "Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA." BuzzFeed News, February 27, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement.

Marcus, Gary, and Ernest Davis. *Rebooting AI*. Penguin Random House, 2019. https://www.penguinrandomhouse.com/books/603982/rebooting-ai-by-gary-marcus-and-ernest-davis/9781524748258.

Miller, Matthew Thomas, Maxim G. Romanov, and Sarah Bowen Savant. "Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans." *International Journal of Middle East Studies* 50, no. 1 (February 2018): 103–9. https://doi.org/10.1017/S0020743817000964.

Milligan, Ian. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." *The Canadian Historical Review* 94, no. 4 (2013): 540–69.

Mishra, Vidisha, and Madhulika Srikumar. "Predatory Data: Gender Bias in Artificial Intelligence." *CyFy Journal*, 2017, 67–71.

MIT Work of the Future. "The Work of the Future: Shaping Technology and Institutions," Fall 2019.

Mittal, Ankush, and Sumit Gupta. "Automatic Content-Based Retrieval and Semantic Classification of Video Content." *International Journal on Digital Libraries* 6, no. 1 (February 1, 2006): 30–38. https://doi.org/10.1007/s00799-005-0119-y.

Mühling, Markus, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth, and Bernd Freisleben. "Content-Based Video Retrieval in Historical Collections of the German Broadcasting Archive." *International Journal on Digital Libraries* 20, no. 2 (June 1, 2019): 167–83. https://doi.org/10.1007/s00799-018-0236-z.

Mullen, Lincoln. "America's Public Bible: Biblical Quotations in U.S. Newspapers," 2016. https://americaspublicbible.org/.

Munyaradzi, Ngoni, and Hussein Suleman. "A System for High Quality Crowdsourced Indigenous Language Transcription." *International Journal on Digital Libraries* 14, no. 3 (August 1, 2014): 117–25. https://doi.org/10.1007/s00799-014-0112-4.

Murphy, Oonagh, and Elena Villaespesa. "AI: A Museum Planning Toolkit." Goldsmiths, University of London, January 2020. The Museums + AI Network. https://themuseumsainetwork.files.wordpress.com/2020/02/20190317_museums-and-ai-toolkit_rl_web.pdf.

Nelson, Laura. Machine Learning and Libraries Interview with Laura Nelson. Interview by Ryan Cordell, March 17, 2020.

Nelson, Laura K. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research*, November 21, 2017, 0049124117729703. https://doi.org/10.1177/0049124117729703.

Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. "How We Do Things with Words: Analyzing Text as Social and Cultural Data." *ArXiv:1907.01468 [Cs]*, July 2, 2019. http://arxiv.org/abs/1907.01468.

Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism*. 1 edition. New York: NYU Press, 2018.

Noble, Safiya Umoja. "Safiya Umoja Noble PhD on Twitter." Twitter, June 2, 2020. https://twitter.com/safiyanoble/status/1267978038072889344.

O'Connell, Pamela LiCalzi. "One Picture, 1,000 Tags." *The New York Times*, March 28, 2007, sec. Arts. https://www.nytimes.com/2007/03/28/arts/artsspecial/28social.html.

Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. "DhSegment: A Generic Deep-Learning Approach for Document Segmentation." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, August 2018, 7–12. https://doi.org/10.1109/ICFHR-2018.2018.00011.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1 edition. New York: Crown, 2016.

Ọnụọha, Mimi. "MIMI ỌNỤỌHA." MIMI ỌNỤỌHA. Accessed February 20, 2020. http://mimionuoha.com.

Padilla, Thomas. "On a Collections as Data Imperative." Accessed February 4, 2020. http://digitalpreservation.gov/meetings/dcs16/tpadilla_OnaCollectionsasDataImperative_final.pdf.

———. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." OCLC RESEARCH POSITION PAPER. Dublin, Ohio: OCLC Research, December 9, 2019. https://www.oclc.org/content/dam/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.pdf.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: Sentiment Classification Using Machine Learning Techniques." In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02*, 10:79–86. Not Known: Association for Computational Linguistics, 2002. https://doi.org/10.3115/1118693.1118704.

Park, Jung-ran, and Caimei Lu. "Application of Semi-Automatic Metadata Generation in Libraries: Types, Tools, and Techniques." *Library & Information Science Research* 31, no. 4 (December 1, 2009): 225–31. https://doi.org/10.1016/j.lisr.2009.05.002.

Pearl, Lisa, and Mark Steyvers. "Detecting Authorship Deception: A Supervised Machine Learning Approach Using Author Writeprints." *Literary and Linguistic Computing* 27, no. 2 (June 1, 2012): 183–96. https://doi.org/10.1093/llc/fqs003.

Perez, Angelica. "EmoPy: A Machine Learning Toolkit for Emotional Expression." ThoughtWorks, August 24, 2018. https://www.thoughtworks.com/insights/blog/emopy-machine-learning-toolkit-emotional-expression.

Peters, Jay. "IBM Will No Longer Offer, Develop, or Research Facial Recognition Technology." The Verge, June 8, 2020. https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-recognition-analysis-software.

Piatetsky, Gregory. "20 AI, Data Science, Machine Learning Terms You Need to Know in 2020 (Part 2)." *KDnuggets* (blog), March 2020. https://www.kdnuggets.com/20-ai-data-science-machine-learning-terms-you-need-to-know-in-2020-part-2.html/.

"Picture What the Cloud Can Do | Google Cloud." Accessed September 17, 2019. https://cloud.withgoogle.com/nytimes/.

Potter, Abigail. "Report Draft Comments," May 15, 2020.

Berkeley Institute for Data Science. "Public Editor," July 31, 2019. https://bids.berkeley.edu/research/public-editor.

Race, Technology, and Algorithmic Bias | Vision & Justice, 2019. https://www.radcliffe.harvard.edu/video/race-technology-and-algorithmic-bias-vision-justice.

Radford, Alec, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. "Better Language Models and Their Implications." OpenAI, February 14, 2019. https:

//openai.com/blog/better-language-models/.

Reiff Conell, Sarah, Lingdong Huang, Golan Levin, and Matthew D. Lincoln. "National Neighbors - CMU DH." Accessed November 6, 2019. https://dh-web.hss.cmu.edu/nga/essay.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability." AI Now Institute, April 2018.

Resig, John. "Ukiyo-e.Org: Aggregating and Analyzing Digitized Japanese Woodblock Prints." Vimeo, September 16, 2013. https://vimeo.com/74691102.

Richardson, Rashida, Jason M. Schultz, and Kate Crawford. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice." *New York University Law Review* 94, no. 192 (May 2019). /reports.html.

Ridge, Mia. "Museums + AI, New York Workshop Notes – Open Objects," September 18, 2019. http://www.openobjects.org.uk/2019/09/museums-ai-new-york-workshop-notes/.

Ridley, Michael. "Explainable Artificial Intelligence (RLI 299, 2019)." *Research Library Issues*, no. 299 (2019): 28–46. https://doi.org/10.29242/rli.299.3.

Rockwell, Geoffrey, and Bettina Berendth. "Information Wants to Be Free, Or Does It?: The Ethics of Datafication | Electronic Book Review." *Electronic Book Review*, December 3, 2017. https://electronicbookreview.com/essay/information-wants-to-be-free-or-does-it-the-ethics-of-datafication/.

Rust, Amanda. Machine Learning and Libraries Interview with Amanda Rust. Interview by Ryan Cordell, April 29, 2020.

Sandler, Ronald, John Basl, and Steven C. Tiell. "Building Data and AI Ethics Committees | Accenture," August 20, 2019. https://www.accenture.com/us-en/insights/software-platforms/building-data-ai-ethics-committees.

Schmidt, Ben. "Sapping Attention: Machine Learning at Sea." *Sapping Attention* (blog), November 1, 2012. http://sappingattention.blogspot.com/2012/11/machine-learning-on-high-seas.html.

Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI." *ArXiv:1907.10597 [Cs, Stat]*, August 13, 2019. http://arxiv.org/abs/1907.10597.

SebastianiFabrizio. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys (CSUR)*, March 1, 2002. https://dl.acm.org/doi/abs/10.1145/505282.505283.

Shahaf, Dafna, Carlos Guestrin, and Eric Horvitz. "Trains of Thought: Generating Information Maps." In *Proceedings of the 21st International Conference on World Wide Web - WWW '12*, 899–908. Lyon, France: ACM Press, 2012. https://doi.org/10.1145/2187836.2187957.

Shane, Janelle. "AI Weirdness." AI Weirdness, 2020. https://aiweirdness.com.

Sharp, Oscar. *Sunspring*, 2016. https://www.youtube.com/watch?v=LY7x2Ihqjmc.

Short, Matthew. "Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels." *Cataloging & Classification*

Quarterly 57, no. 5 (July 4, 2019): 315–36. https://doi.org/10.1080/01639374.2019.1653413.

Siguenza-Guzman, Lorena, Victor Saquicela, Elina Avila-Ordóñez, Joos Vandewalle, and Dirk Cattrysse. "Literature Review of Data Mining Applications in Academic Libraries." *The Journal of Academic Librarianship* 41, no. 4 (July 1, 2015): 499–510. https://doi.org/10.1016/j.acalib.2015.06.007.

Sloan, Robin. "Notes from the Quest Factory." Notes from the quest factory, 2019. https://desert.glass/archive/notes-from-the-quest-factory/.

Smith, David A, and Ryan Cordell. "A Research Agenda for Historical and Multilingual Optical Character Recognition," 2018. https://repository.library.northeastern.edu/files/neu:f1881m035.

Smith, Linda C. "Artificial Intelligence in Information Retrieval Systems." *Information Processing & Management* 12, no. 3 (January 1976): 189–222. https://doi.org/10.1016/0306-4573(76)90005-4.

Stack, John. "What the Machine Saw," September 2019. https://johnstack.github.io/what-the-machine-saw/.

"Steve.Museum." In *Wikipedia*, March 6, 2019. https://en.wikipedia.org/w/index.php?title=Steve.museum&oldid=886480771.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." *ArXiv:1906.02243 [Cs]*, June 5, 2019. http://arxiv.org/abs/1906.02243.

Suominen, Osma. "Annif: DIY Automated Subject Indexing Using Multiple Algorithms." *Liber Quarterly* 29, no. 1 (2019). https://doi.org/10.18352/lq.10285.

Taneja, Hemant. "The Era of 'Move Fast and Break Things' Is Over." *Harvard Business Review*, January 22, 2019. https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over.

Tay, Aaron. "How Libraries Might Change When AI, Machine Learning, Open Data, Block Chain & Other Technologies Are the Norm." *Musings about Librarianship* (blog). Accessed November 1, 2019. http://musingsaboutlibrarianship.blogspot.com/2017/04/how-libraries-might-change-when-ai.html.

Tech, Normcore. "Neural Nets Are Just People All the Way Down." Accessed December 19, 2019. https://vicki.substack.com/p/neural-nets-are-just-people-all-the.

The Alan Turing Institute. "Living with Machines." Accessed September 20, 2019. https://www.turing.ac.uk/research/research-projects/living-machines.

The Future of Work in a World of AI, ML, and Automation. MIT Sloan CIO Symposium Videos, 2018. https://www.youtube.com/watch?v=Zg4HfPPx4qo.

The Museums + AI Network. "The Museums + AI Network," February 11, 2020. https://themuseumsai.network/toolkit/.

Thomas, Rachel. "16 Things You Can Do to Make Tech More Ethical, Part 1." fast.ai, April 22, 2019. https://www.fast.ai/2019/04/22/ethics-action-1/.

———. "16 Things You Can Do to Make Tech More Ethical, Part 2." fast.ai, April 25, 2019. https://www.fast.ai/2019/04/25/ethics-action-2/.

———. "16 Things You Can Do to Make Tech More Ethical, Part 3." fast.ai, May 3, 2019. https://www.fast.ai/2019/05/03/ethics-action-3/.

Tilton, Lauren, and Taylor Arnold. Machine Learning and Libraries Interview with Lauren Tilton and Taylor Arnold. Interview by Ryan Cordell, March 3, 2020.

TRANSKRIBUS Team. *Transkribus*. Innsbruck, Austria: University of Innsbruck, 2020. https://transkribus.eu/Transkribus/#archive-content.

Trant, J., and with the participants in the steve. "Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept." *New Review of Hypermedia and Multimedia* 12, no. 1 (June 2006): 83–105. https://doi.org/10.1080/13614560600802940.

Trove. "Text Correction Guidelines." Accessed May 21, 2020. https://help.nla.gov.au/trove/digitised-newspapers/text-correction-guidelines.

Tuarob, Suppawong, Line C. Pouchard, Prasenjit Mitra, and C. Lee Giles. "A Generalized Topic Modeling Approach for Automatic Document Annotation." *International Journal on Digital Libraries* 16, no. 2 (June 1, 2015): 111–28. https://doi.org/10.1007/s00799-015-0146-2.

Underwood, Ted. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127, no. 1 (2014): 64–72.

———. "Why an Age of Machine Learning Needs the Humanities." *Public Books* (blog), December 5, 2018. https://www.publicbooks.org/why-an-age-of-machine-learning-needs-the-humanities/.

Vallor, Shannon, Brian Green, and Irina Raicu. "Ethics in Technology Practice." The Markkula Center for Applied Ethics at Santa Clara University, n.d. https://www.scu.edu/ethics/.

Wagstaff, Kiri L., and Geoffrey Z. Liu. "Automated Classification to Improve the Efficiency of Weeding Library Collections." *The Journal of Academic Librarianship* 44, no. 2 (March 1, 2018): 238–47. https://doi.org/10.1016/j.acalib.2018.02.001.

Walker, Kevin W., and Zhehan Jiang. "Application of Adaptive Boosting (AdaBoost) in Demand-Driven Acquisition (DDA) Prediction: A Machine-Learning Approach." *The Journal of Academic Librarianship* 45, no. 3 (May 1, 2019): 203–12. https://doi.org/10.1016/j.acalib.2019.02.013.

Wang, James Z., Kurt Grieb, Ya Zhang, Ching-chih Chen, Yixin Chen, and Jia Li. "Machine Annotation and Retrieval for Digital Imagery of Historical Materials." *International Journal on Digital Libraries* 6, no. 1 (February 1, 2006): 18–29. https://doi.org/10.1007/s00799-005-0121-4.

West, Sarah Myers, Meredith Whittaker, and Kate Crawford. "Discriminating Systems: Gender, Race, and Power in AI." AI Now Institute, April 2019. https://ainowinstitute.org/discriminatingsystems.html.

White, Philip B. "Using Data Mining for Citation Analysis | White | College & Research Libraries." *College & Research Libraries* 80, no. 1 (2019). https://crl.acrl.org/index.php/crl/article/view/16892.

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schultz. "AI Now Report 2018." AI Now Institute, December 2018.

Willi, Marco, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson. "Identifying Animal Species in Camera Trap Images Using Deep Learning and Citizen Science." *Methods in Ecology and Evolution* 10, no. 1 (2019): 80–91. https://doi.org/10.1111/2041-210X.13099.

Wu, Jian. "Web Science and Digital Libraries Research Group: 2019-10-01: Attending the Machine Learning + Libraries Summit at the Library of Congress." *Web Science and Digital Libraries Research Group* (blog), October 1, 2019. https://ws-dl.blogspot.com/2019/10/2019-10-01-attending-machine-learning.html.

Yale Digital Humanities Lab. "Neural Neighbors." Accessed September 17, 2019. https://dhlab.yale.edu/neural-neighbors/.

———. "PixPlot: Visualizing Image Fields." Accessed September 17, 2019. https://s3-us-west-2.amazonaws.com/lab-apps/pix-plot/index.html?utm_source=dancohen&utm_medium=email.

———. "Yale DHLab - PixPlot." Accessed September 20, 2019. https://dhlab.yale.edu/projects/pixplot/.

Yousif, Hayder, Jianhe Yuan, Roland Kays, and Zhihai He. "Animal Scanner: Software for Classifying Humans, Animals, and Empty Frames in Camera Trap Images." *Ecology and Evolution* 9, no. 4 (2019): 1578–89. https://doi.org/10.1002/ece3.4747.

Zakrzewski, Cat. "Mark Zuckerberg Spoke with Civil Rights Leaders about Trump's Posts. It Didn't Go Well." *Washington Post*, June 2, 2020, sec. PowerPost Analysis Analysis Interpretation of the news based on evidence, including data, as well as anticipating how events might unfold based on past events. https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/02/the-technology-202-mark-zuckerberg-spoke-with-civil-rights-leaders-about-trump-s-posts-it-didn-t-go-well/5ed55da4602ff12947e81457/.

Ziv, Ravid. *Optimization Process in the Information Plane for DNNs*, 2017. https://www.youtube.com/watch?v=q45lPv9rev0&feature=youtu.be.

Zwaard, Kate. Machine Learning and Libraries Interview with Kate Zwaard. Interview by Ryan Cordell, May 27, 2020.

Appendix B: Interviewees

These interviews were not intended as an exhaustive survey of the fields drawn from in preparation of this report. These interviews were conceived as supplements to the presentations given at the Machine Learning + Libraries summit at the Library of Congress in September 2019, as well as of the literature review, which aimed to understand the much broader ML and libraries conversation. Each interview lasted 30-60 minutes.

Name	Title	Interview Date
Thomas Padilla	Interim Head, Knowledge Production, University of Nevada Las Vegas	February 4, 2020

Name	Title	Interview Date
Benjamin Charles Germain Lee	Innovator-in-Residence, Library of Congress and Ph.D. Student, Computer Science, University of Washington	February 6, 2020
Ted Underwood	Professor, Information Sciences and English, University of Illinois	February 6, 2020
Dan Cohen	Dean of Libraries, Vice Provost for Information Collaboration, and Professor of History, Northeastern University	February 7, 2020
Elizabeth Lorang	Associate Professor, Libraries, University of Nebraska—Lincoln	February 21, 2020
Lauren Tilton (with Taylor Arnold)	Assistant Professor of Digital Humanities, University of Richmond	March 3, 2020
Taylor Arnold (with Lauren Tilton)	Assistant Professor of Statistics, University of Richmond	March 3, 2020
Laura Nelson	Assistant Professor, Sociology, Northeastern University	March 17, 2020
Amanda Rust	Associate Director, Digital Scholarship Group and Digital Humanities Librarian, Northeastern University	April 28, 2020
Kate Zwaard	Director of Digital Strategy, Library of Congress	May 27, 2020

Appendix C: ML + Libraries Meeting Presenters

The following people presented at the Museums + Libraries Summit at the Library of Congress on September 20, 2019 and contributed their slides to the meeting documentation. Their presentations and contributions to the summit's discussions, group activities, and artifacts influence nearly every aspect of this report.

Name	Title and Affiliation
Nick Adams	Chief Scientist at the Goodly Labs

Name	Title and Affiliation
Audrey Altman	Developer, Digital Public Library of America
Shawn Averkamp	Senior Consultant, AVP
John Bell	Digital Humanities Program Manager, Dartmouth College Research Computing
Karen Cariani	Executive Director, WGBH Media Library and Archives
Tom Cramer	Chief Technology Strategist, Stanford Libraries
Hannah Davis	Independent Research Artist
Jon Dunn	Assistant Dean for Library Technologies, Indiana University Bloomington
Rebecca Dikow	Research Data Scientist, Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution
Corey DiPietro	Collections Information Manager, National Museum of American History
Michael Haley Goldman	US Holocaust Memorial Museum
Ross Goodwin	Data Artist
Josh Hadro	Managing Director, IIIF Consortium
John Hessler	Library of Congress, Curator of the Jay I. Kislak collection of the Archaeology & History of the Early Americas; Specialist in Computational Geography & Geographic Information Science
Benjamin Charles Germain Lee	Ph.D. Student, University of Washington Computer Science and Engineering
Peter Leonard	Director, Yale DH Lab
Michael Lesk	Professor, Rutgers University
Kurt Luther	Assistant Professor of Computer Science and (by courtesy) History, Virginia Tech
Harish Maringanti	Associate Dean for IT & Digital Library Services; J. Willard Marriott Library at University of Utah
Anishi Mehta	Computer Science student, Georgia Institute of Technology; LC Labs Innovation Intern, Summer 2019
Thu Phuong 'Lisa' Nguyen	Curator for Digital Scholarship & Asian Initiatives, Hoover Institution Library & Archives, Stanford University
Mia Ridge	Digital Curator, British Library
Helena Sarin	Neural Bricolage, Founder
David Smith	Associate Professor, Computer Science, Northeastern University

Name	Title and Affiliation
Leen-Kiat Soh	Professor of Computer Science, University of Nebraska-Lincoln
Mike Trizna	Data Scientist, Smithsonian Institution
Mark Williams	Associate Professor of Film and Media Studies, Dartmouth College
Heather Yager	Associate Director for Technology, MIT Libraries

Appendix D: ML + Libraries Summit Event Summary

The ML + Libraries Event Summary can be found at https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf and may be appended directly to PDF versions of this report.

Works Cited

Abebe, Rediet, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. "Roles for Computing in Social Change," January 28, 2020. https://doi.org/10.1145/3351095.3372871.

AI Now. "Algorithmic Accountability Policy Toolkit," October 2018.

Algorithmic Justice League. "Safe Face Pledge." Center on Privacy & Technology at Georgetown Law, 2018. https://www.safefacepledge.org/.

Alpert-Abrams, Hannah. "Machine Reading the Primeros Libros" 10, no. 4 (2016). http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html.

Amnesty International, and Access Now. "The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems," May 16, 2018. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

Amodei, Dario, Danny Hernandez, Girish Sastry, Jack Clark, Greg Brockman, and Ilya Sutskever. "AI and Compute." OpenAI, November 7, 2019. https://openai.com/blog/ai-and-compute/.

Ardanuy, Mariona Coll, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. "Living Machines: A Study of Atypical Animacy," May 22, 2020. http://arxiv.org/abs/2005.11140.

Arnold, Taylor, Lauren Tilton, and Annie Berke. "Visual Style in Two Network Era Sitcoms." *Journal of Cultural Analytics*, 2019. https://doi.org/10.22148/16.043.

Basaran, Dogac, Slim Essid, and Geoffroy Peeters. "MAIN MELODY EXTRACTION WITH SOURCE-FILTER NMF AND CRNN." In *19th International Society for Music Information Retreival*. Paris, France, 2018. https://hal.archives-ouvertes.fr/hal-02019103.

Beals, M. H., and Emily Bell. "The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges." Loughborough, 2020. https://www.digitisednewspapers.net/.

Bourg, Chris. "What Happens to Libraries and Librarians When Machines Can Read All the Books?" Feral Librarian, March 17, 2017. https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/.

Brygfjeld, Svein Arne, Freddy Wetjen, and André Walsøe. "Machine Learning for Production of Dewey Decimal," July 19, 2018, 9.

Buolamwini, Joy. "The Coded Gaze." AJL -ALGORITHMIC JUSTICE LEAGUE, November 6, 2016. https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148.

Byrum, Joseph. "Build a Diverse Team to Solve the AI Riddle." MIT Sloan Management Review, May 18, 2020. https://sloanreview.mit.edu/article/build-a-diverse-team-to-solve-the-ai-riddle/.

Caplan, Robyn, Lauren Hanson, and Jeanna Matthews. "Algorithmic Accountability: A Primer." Data & Society. Accessed January 28, 2020. https://datasociety.net/output/algorithmic-accountability-a-primer/.

Chen, Hsinchun, Terence R. Smith, Mary L. Larsgaard, Linda L. Hill, and Marshall Ramsey. "A Geographic Knowledge Representation System for Multimedia Geospatial Retrieval and Analysis." *International Journal on Digital Libraries* 1, no. 2 (September 1, 1997): 132–52. https://doi.org/10.1007/s007990050010.

Congress, Library of. "Beyond Words," 2020. http://beyondwords.labs.loc.gov/#/.

Cordell, Ryan. "Machine Learning and Libraries Interview with Amanda Rust," April 29, 2020.

———. "Machine Learning and Libraries Interview with Benjamin Charles Germain Lee," February 6, 2020.

———. "Machine Learning and Libraries Interview with Dan Cohen," February 7, 2020.

———. "Machine Learning and Libraries Interview with Elizabeth Lorang," February 21, 2020.

———. "Machine Learning and Libraries Interview with Kate Zwaard," May 27, 2020.

———. "Machine Learning and Libraries Interview with Laura Nelson," March 17, 2020.

———. "Machine Learning and Libraries Interview with Lauren Tilton and Taylor Arnold," March 3, 2020.

———. ""Q I-Jtb the Raven": Taking Dirty OCR Seriously." *Book History* 20 (2017): 188–225.

Cordell, Ryan, and David Smith. "Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines." Digital History Project. The Viral Texts Project, 2020. https://viraltexts.org.

Crawford, Kate, and Trevor Paglen. "Excavating AI: The Politics of Training Sets for Machine Learning." -, September 19, 2019. https://www.excavating.ai.

Critical Platform Studies Group. "Algorithmic Equity Toolkit." ACLU of Washington, April 2020. https://www.aclu-wa.org/AEKit.

Davis, Hannah. "TransProse," 2013. http://www.musicfromtext.com/.

Dayma, Boris. "HuggingTweets - Train a Model to Generate Tweets." W&B, 2020. https://tinyurl.com/ybaqmhzo.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Cambridge, Massachusetts: The MIT Press, 2020.

"Dig That Lick," 2019. http://dig-that-lick.eecs.qmul.ac.uk/Dig%20That%20Lick_About.html.

Earl, Chandra, Alexander White, Michael Trizna, Paul Frandsen, Akito Kawahara, Sean Brady, and Rebecca Dikow. "Discovering Patterns of Biodiversity in Insects Using Deep Machine Learning." *Biodiversity Information Science and Standards* 3 (July 2, 2019): e37525. https://doi.org/10.3897/biss.3.37525.

Esposito, Floriana, Donato Malerba, Giovanni Semeraro, Nicola Fanizzi, and Stefano Ferilli. "Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service." *International Journal on Digital Libraries* 2, no. 1 (October 1, 1998): 3–19. https://doi.org/10.1007/s007990050033.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.

Fagan, Benjamin. "Chronicling White America." *American Periodicals: A Journal of History & Criticism* 26, no. 1 (2016): 10–13. https://muse.jhu.edu/journals/american_periodicals/v026/26.1.fagan.html.

Ferris, Anna M. "The Ethics and Integrity of Cataloging." *Journal of Library Administration* 47, nos. 3-4 (July 1, 2008): 173–90. https://doi.org/10.1080/01930820802186514.

Ferriter, Meghan. "Report Draft Comments," May 15, 2020.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." *SSRN Electronic Journal*, 2020. https://doi.org/10.2139/ssrn.3518482.

Gallagher, Liz. "Finding Image Pathways." Medium. Accessed September 17, 2019. https://stacks.wellcomecollection.org/finding-image-pathways-12d31ae347f9.

Gotterbarn, Don, Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Vazansky, Marty J Wolf, et al. "ACM Code of Ethics and Professional Conduct," 2018, 28.

Horton, Russell, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie." *Digital Humanities Quarterly* 003, no. 2 (June 18, 2009).

IBM. "AI Explainability 360 Open Source Toolkit." GitHub, 2019. http://aix360.mybluemix.net/.

"Image Analysis for Archival Discovery (Aida)," 2020. http://projectaida.org/.

Jakeway, Eileen. "Newspaper Navigator Surfaces Treasure Trove of Historic Images a Sneak Peek at Upcoming Data Jam! | the Signal." Webpage, April 21, 2020. https://blogs.loc.gov/thesignal/2020/04/newspaper-navigator-surfaces-treasure-trove-of-historic-images-get-a-sneak-peek-at-upcoming-data-jam/.

Jakeway, Eileen, Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, Abigail Potter, and Kate Zwaard. "Machine Learning + Libraries Summit Event Summary." LC Labs Digital Strategy Directorate, February 13, 2020. https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

Jo, Eun Seo, and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning," December 22, 2019. https://doi.org/10.1145/3351095.3372829.

Katell, Michael, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Binz, Daniella Raz, and P. M. Krafft. "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 45–55. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372874.

Keralis, Spencer D. C. "Milking the Deficit Internship." In *Disrupting the Digital Humanities*, edited by Dorothy Kim and Jesse Stommel, 2016. http://www.disruptingdh.com/milking-the-deficit-internship/.

King, Rachael Scarborough. *Writing to the World: Letters and the Origins of Modern Print Genres*. Baltimore: Johns Hopkins University Press, 2018.

Lee, Benjamin Charles Germain. "Machine Learning, Template Matching, and the International Tracing Service Digital Archive: Automating the Retrieval of Death Certificate Reference Cards from 40 Million Document Scans." *Digital Scholarship in the Humanities* 34, no. 3 (September 1, 2019): 513–35. https://doi.org/10.1093/llc/fqy063.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. "The Newspaper Navigator Dataset: Extracting and Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America," May 4, 2020. http://arxiv.org/abs/2005.01583.

Lee, Maurice S. "Falsifiability, Confirmation Bias, and Textual Promiscuity." *J19: The Journal of Nineteenth-Century Americanists* 2, no. 1 (April 3, 2014): 162–71. https://doi.org/10.1353/jnc.2014.0014.

Library of Congress. "Digital Strategy," April 26, 2019. https://www.loc.gov/digital-strategy.

Lincoln, Matthew D., Golan Levin, Sarah Reiff Conell, and Lingdong Huang. "National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections," November 2019, 20. https://nga-neighbors.library.cmu.edu.

Linder, Courtney. "The Librarians of the Future Will Be AI Archivists." Popular Mechanics, May 13, 2020. https://www.popularmechanics.com/technology/a32436235/library-of-congress-machine-learning-newspaper-images/.

Lorang, Prepared Elizabeth, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project," January 10, 2020, 43.

Miller, Matthew Thomas, Maxim G. Romanov, and Sarah Bowen Savant. "Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans." *International Journal of Middle East Studies* 50, no. 1 (February 2018): 103–9. https://doi.org/10.1017/S0020743817000964.

Milligan, Ian. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 19972010." *The Canadian Historical Review* 94, no. 4 (2013): 540–69. https://muse-jhu-edu.ezproxy.neu.edu/journals/canadian_historical_review/v094/94.4.milligan01.html.

Mullen, Lincoln. "America's Public Bible: Biblical Quotations in U.S. Newspapers," 2016. https://americaspublicbible.org/.

Munyaradzi, Ngoni, and Hussein Suleman. "A System for High Quality Crowdsourced Indigenous Language Transcription." *International Journal on Digital Libraries* 14, no. 3 (August 1, 2014): 117–25. https://doi.org/10.1007/s00799-014-0112-4.

Murphy, Oonagh, and Elena Villaespesa. "AI: A Museum Planning Toolkit." Goldsmiths, University of London, January 2020. https://themuseumsainetwork.files.wordpress.com/2020/02/20190317_museums-and-ai-toolkit_rl_web.pdf.

Mühling, Markus, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth, and Bernd Freisleben. "Content-Based Video Retrieval in Historical Collections of the German

Broadcasting Archive." *International Journal on Digital Libraries* 20, no. 2 (June 1, 2019): 167–83. https://doi.org/10.1007/s00799-018-0236-z.

Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism*. 1 edition. New York: NYU Press, 2018.

Noble, Safiya Umoja. "Safiya Umoja Noble PhD on Twitter." Twitter, June 2, 2020. https://twitter.com/safiyanoble/status/1267978038072889344.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1 edition. New York: Crown, 2016.

Padilla, Thomas. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." OCLC RESEARCH POSITION PAPER. Dublin, Ohio: OCLC Research, December 9, 2019. https://www.oclc.org/content/dam/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.pdf.

Peters, Jay. "IBM Will No Longer Offer, Develop, or Research Facial Recognition Technology." The Verge, June 8, 2020. https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-recognition-analysis-software.

Potter, Abigail. "Report Draft Comments," May 15, 2020.

Radford, Alec, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. "Better Language Models and Their Implications." OpenAI. Accessed November 13, 2019. https://openai.com/blog/better-language-models/.

Reiff Conell, Sarah, Lingdong Huang, Golan Levin, and Matthew D. Lincoln. "National Neighbors - CMU DH." Accessed November 6, 2019. https://dh-web.hss.cmu.edu/nga/essay.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability." AI Now Institute, April 2018.

Ridge, Mia. "Museums + AI, New York Workshop Notes," September 18, 2019. http://www.openobjects.org.uk/2019/09/museums-ai-new-york-workshop-notes/.

Ridley, Michael. "Explainable Artificial Intelligence (RLI 299, 2019)." *Research Library Issues*, no. 299 (2019): 28–46. https://doi.org/10.29242/rli.299.3.

Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI," August 13, 2019. http://arxiv.org/abs/1907.10597.

Shane, Janelle. "AI Weirdness," 2020. https://aiweirdness.com.

Smith, David A, and Ryan Cordell. "A Research Agenda for Historical and Multilingual Optical Character Recognition," 2018. https://repository.library.northeastern.edu/files/neu:f1881m035.

Smith, Linda C. "Artificial Intelligence in Information Retrieval Systems." *Information Processing & Management* 12, no. 3 (January 1976): 189–222. https://doi.org/10.1016/0306-4573(76)90005-4.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP," June 5, 2019. http://arxiv.org/abs/1906.02243.

Sunspring, 2016. https://www.youtube.com/watch?v=LY7x2Ihqjmc.

Taneja, Hemant. "The Era of 'Move Fast and Break Things' Is Over." *Harvard Business Review*, January 22, 2019. https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over.

Thomas, Rachel. "16 Things You Can Do to Make Tech More Ethical, Part 1." fast.ai, April 22, 2019. https://www.fast.ai/2019/04/22/ethics-action-1/.

———. "16 Things You Can Do to Make Tech More Ethical, Part 2." fast.ai, April 25, 2019. https://www.fast.ai/2019/04/25/ethics-action-2/.

TRANSKRIBUS Team. *Transkribus*. Innsbruck, Austria: University of Innsbruck, 2020. https://transkribus.eu/Transkribus/#archive-content.

Trove. "Text Correction Guidelines." Accessed May 21, 2020. https://help.nla.gov.au/trove/digitised-newspapers/text-correction-guidelines.

Underwood, Ted. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127, no. 1 (2014): 64–72. http://rep.ucpress.edu/content/127/1/64.abstract.

———. "Why an Age of Machine Learning Needs the Humanities." Public Books, December 5, 2018. https://www.publicbooks.org/why-an-age-of-machine-learning-needs-the-humanities/.

Wagstaff, Kiri L., and Geoffrey Z. Liu. "Automated Classification to Improve the Efficiency of Weeding Library Collections." *The Journal of Academic Librarianship* 44, no. 2 (March 1, 2018): 238–47. https://doi.org/10.1016/j.acalib.2018.02.001.

Yale Digital Humanities Lab. "Neural Neighbors." Accessed September 17, 2019. https://dhlab.yale.edu/neural-neighbors/.

———. "PixPlot: Visualizing Image Fields." Accessed September 17, 2019. https://s3-us-west-2.amazonaws.com/lab-apps/pix-plot/index.html?utm_source=dancohen&utm_medium=email.

Zakrzewski, Cat. "Mark Zuckerberg Spoke with Civil Rights Leaders About Trump's Posts. It Didn't Go Well." *Washington Post: PowerPost Analysis Analysis Interpretation of the News Based on Evidence, Including Data, as Well as Anticipating How Events Might Unfold Based on Past Events*. Accessed June 3, 2020. https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/02/the-technology-202-mark-zuckerberg-spoke-with-civil-rights-leaders-about-trump-s-posts-it-didn-t-go-well/5ed55da4602ff12947e81457/.