Responsible Operations: Data Science, Machine Learning, and AI in Libraries

Thomas Padilla
Practitioner Researcher in Residence

OCLC

December 2019

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-152-1
DOI: 10.25333/xk7z-9g97
OCLC Control Number: 1129383585

ORCID iDs
Thomas Padilla, https://orcid.org/0000-0002-6743-6592

Please direct correspondence to:
OCLC Research
oclcresearch@oclc.org

Suggested citation:
Padilla, Thomas. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. https://doi.org/10.25333/xk7z-9g97.

We would understand that the strength of our movement is in the strength of our relationships, which could only be measured by their depth. Scaling up would mean going deeper, being more vulnerable and more empathetic.

—adrienne maree brown[1]

CONTENTS

INTRODUCTION

In light of widespread interest, OCLC commissioned the development of a research agenda to help chart library community engagement with data science, machine learning, and artificial intelligence (AI).[2] *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* is the result.[3] *Responsible Operations* was developed in partnership with an advisory group and a landscape group from March 2019 through September 2019. The suggested areas of investigation represented in this agenda present interdependent technical, organizational, and social challenges. Given the scale of positive and negative impacts presented by the use of data science, machine learning, and AI, addressing these challenges together is more important than ever.[4]

Consider this agenda multifaceted. Rather than presenting the technical, organizational, and social challenges as separate considerations, each new facet is integral to the structure of the whole. For example, when prompted to provide challenges and opportunities for this agenda, multiple contributors suggested that libraries could use machine learning to increase access to their collections. When this suggestion was shared with other contributors, they affirmed it while turning it on its axis, revealing additional planes of engagement.

All agreed that the challenge of doing this work responsibly requires fostering organizational capacities for critical engagement, managing bias, and mitigating potential harm.

Some contributors suggested that the challenge for this work resides in the community coming to a better understanding of the staff competencies, experiences, and dispositions that support utilization and development of machine learning in a library context. Other contributors focused on the evergreen organizational challenge of moving emerging work from the periphery to the core. All agreed that the challenge of doing this work responsibly requires fostering organizational capacities for critical engagement, managing bias, and mitigating potential harm.

Accordingly, the 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 interdisciplinary collaboration." All the above has no foundation without "committing to responsible operations."

Responsible Operations lays groundwork for this commitment: it is a call for action.

No single country, association, or organization can meet the challenges that lie ahead. Progress will benefit from diverse collaborations forged among librarians, archivists, museum professionals, computer scientists, data scientists, sociologists, historians, human computer

interaction experts, and more. All have a role to play. For this reason, and at this stage, we have framed the agenda in terms of general recommendations for action, without specifying particular agencies in each case. We hope that particular groups will be encouraged to take action by the recommendations here. And certainly, OCLC looks forward to partnering on, contributing to, and amplifying efforts in this field.

AUDIENCE

- **Library administrators, faculty, and staff** who are interested in engaging on core challenges with data science, machine learning, and AI. Challenges for this audience are necessarily technical, organizational, and social in nature.
- **University administrators and disciplinary faculty** who want to collaborate with library administrators, faculty, and staff on challenges that strike a balance between research, pedagogy, and professional practice.
- **Professionals** operating in commercial and nonprofit contexts interested in collaborating with library administrators, faculty, and staff. The commercial audience in particular has an opportunity to make technology, methods, and data less opaque. Greater transparency would help foster a more collaborative environment.
- **Funders** invested in supporting sustainable, ethically grounded innovation in libraries and cultural heritage organizations more generally.

PROCESS AND SCOPE

Responsible Operations is the product of synchronous development that ran March 2019 through August 2019. These engagements included meetings with an advisory group and interviews and a face-to-face event with a "landscape" group. The landscape group is composed of individuals working in libraries and a group of external experts.[5] Library staff were selected with an eye toward diversification of roles and institutional affiliations. Asynchronous development of the agenda continued August 2019 through September 2019, during which, the author sought feedback and further contributions from the advisory group and the landscape group. Given time constraints for development, the agenda primarily represents the perspective of individuals working in the United States—future efforts will expand to engage international communities.

GUIDING PRINCIPLE

This agenda adapts Rumman Chowdhury's concept of **responsible operations** as a guiding principle.[6] In the context of this agenda, responsible operations refers to individual, organizational, and community capacities to support responsible use of data science, machine learning, and AI. While the principle is not always explicitly stated, all suggested areas of investigation are guided by it.

Throughout the course of agenda development, contributors expressed concern that an increased adoption of algorithmic methods could lead to the amplification of bias that negatively impacts library staff, library users, and society more broadly. Contributors nearly uniformly agreed that the library community shows growing awareness of a topic like algorithmic bias—a path paved to some extent by preexisting work on bias in description and collection development.[7] The work of scholars like Safiya Noble and subsequent activity of practitioners like Jason Clark have extended these areas of library consideration, creating spaces for critical discussion of algorithmic influence in daily life.[8]

Despite greater awareness, significant gaps persist between concept and operationalization in libraries at the level of workflows (managing bias in probabilistic description), policies (community engagement vis-à-vis the development of machine-actionable collections), positions (developing staff who can utilize, develop, critique, and/or promote services

influenced by data science, machine learning, and AI), collections (development of "gold standard" training data), and infrastructure (development of systems that make use of these technologies and methods). Shifting from awareness to operationalization will require holistic organizational commitment to responsible operations. The viability of responsible operations depends on organizational incentives and protections that promote constructive dissent.

> Successful governance of AI systems need to allow "constructive dissent"—that is, a culture where individuals, from the bottom up, are empowered to speak and protected if they do so. It is self-defeating to create rules of ethical use without the institutional incentives and protections for workers engaged in these projects to speak up. (Rumman Chowdhury)[9]

Despite greater awareness, significant gaps persist between concept and operationalization in libraries at the level of workflows, policies, positions, collections, and infrastructure.

Responsible operations and constructive dissent are grounded by shared ethical commitments. As libraries seek to evaluate the extent to which existing ethical commitments account for the positive and negative effects of algorithmic methods, they would do well to engage with Luciano Floridi's and Joshua Cowls's "A Unified Framework of Five Principles for AI in Society." The framework provides five overarching principles to guide the development and use of artificial intelligence:

- **beneficence**—promoting well-being, preserving dignity, and sustaining the planet
- **nonmaleficence**—privacy, security, and "capability caution"
- **autonomy**—the power to decide
- **justice**—promoting prosperity, preserving solidarity, and avoiding unfairness
- **explicability**—enabling the other principles through intelligibility and accountability[10]

No discussion of technology or ethics is complete without critical historical context (e.g., algorithmic bias as a phenomenon with historical precedent and the potential to negatively impact marginalized communities).[11] Refracting potential library effort through these lenses can only serve to increase the efficacy of responsible operations.

Areas of Investigation

The following areas of investigation consist of seven high-level categories paired with subsets of challenges and recommendations. Recommendations are not prioritized, nor is leadership for community investigation designated. This is purposeful given the relative maturity of work and a desire that a range of leadership will advance approaches for addressing challenges. Notions of community are broad, geared toward accommodating action within and across professional and disciplinary communities.

Areas of investigation are interdependent in nature, calling for synchronous work across the seven areas:

1. Committing to Responsible Operations
2. Description and Discovery
3. Shared Methods and Data
4. Machine-Actionable Collections
5. Workforce Development
6. Data Science Services
7. Sustaining Interprofessional and Interdisciplinary Collaboration

COMMITTING TO RESPONSIBLE OPERATIONS

Libraries express increased interest in the use of algorithmic methods. The reasons are many and include, but are not limited to, creating efficiencies in collection description, discovery, and access; freeing up staff time to meet evolving demands; and improving the overall library user experience. Concerns about the potential negative impacts of these methods are significant, and concrete use cases are readily available for libraries to consider. For example, facial recognition technology has been applied to historic cultural heritage collections in a manner that works to increase agency for historically marginalized populations.[12] Yet, use of a technology like this must be weighed relative to a broader field of misuse spanning applications that lack the ability to recognize the faces of people of color, that discriminate based on color, and that foster a capacity for discrimination based on sexuality. In cases like these, the balance of evaluation may result in a decision to not use a technology—a positive outcome for responsible operations.[13] By committing to responsible operations, the library community works to do good with data science, machine learning, and AI.

Managing Bias

Responsible operations call for sustained engagement with human biases manifest in training data, machine learning models, and outputs. In contrast to some discussions that frame algorithmic bias or bias in data as something that can be eliminated, Nicole Coleman suggests that libraries might be better served by focusing on approaches to **managing bias**.[14] 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. Of course, this is not an ahistorical frame of thinking. After all, many areas of library practice find themselves working to address harms their practices have posed, and continue to pose, for marginalized communities.[15] As libraries seek to improve bias-management activities, progress will be continually limited by lack of diversity in staffing; monoculture cannot effectively manage bias.[16] Diversity is not an option, it is an imperative.

Recommendations:

1. 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.

2. Explore creation of a "practices exchange" that highlights successes as well as notable missteps in cultural heritage use of data science, machine learning, and AI. Commit to transparency as a means to work against repeated community mistakes—a pattern of negative behavior in Silicon Valley that Jacob Metcalf, Emanuel Moss, and danah boyd have referred to as "blinkered isomorphism."[17]
3. Synthesize existing guidance on formation of committees (e.g., scope, power, roles, and diversification of identity, experience, and expertise) that can help guide responsible engagement with machine learning and artificial intelligence. Assess to what degree modifications need to be made particular to library contexts.[18]
4. Develop a range of approaches to auditing data science, machine learning, and AI approaches (e.g., the product of computer vision applied to a collection, the strengths and weaknesses of training data on which a model was trained).
5. Convene pilots and working groups focused on adapting and operationalizing work surfaced by recommendations 1–4.

Transparency, Explainability, and Accountability

Responsible operational use of data science, machine learning, and AI depends on making the design, function, and intent of these approaches as transparent as possible. Per efforts by entities like IBM and Microsoft, transparency must go hand in hand with practices that encourage "explainability."[19] Research agenda contributors called for practices and systems that facilitate user interaction with data on both ends of an algorithm—the data that goes in and the data that goes out. For example, libraries might implement a publicly accessible "track changes" feature for data description so any changes to a collection—whether the product of direct human description or algorithmic probability—can be viewed and are machine-actionable.[20]

Responsible operations that embed transparency and explainability increase the likelihood of organizational accountability.

Libraries might also work to support the development of public-facing collection interfaces that allow a user to adjust the parameters of an algorithm being applied to a collection, foregrounding rather than occluding the subjective nature of collection representation. Jer Thorp recommended potential use of a version control system, akin to Git, that would allow users to raise issues with a collection, like biasing out of step with proposed ethical values.[21] Relatedly, a system in this vein might help facilitate recording the provenance of training data—a pervasive lack that compromises potential applications. Responsible operations that embed transparency and explainability increase the likelihood of organizational accountability.

Recommendations:

1. Form a working group focused on documenting efforts to make machine learning and artificial intelligence more transparent and explainable. Synthesize findings into proposed best practices.
2. Conduct a range of pilots that seek to make machine learning and artificial intelligence more transparent and explainable to library staff and library users—be guided by outcomes of recommendation 1.
3. Evaluate practices and systems that encourage organizational accountability in the context of algorithmic work. Share and propose potential next steps.

4. Evaluate terms and conditions associated with cloud-based, off-the-shelf machine learning and AI solutions. Assess the degree to which terms and conditions and/or licensing agreements (a) breach expectations of privacy and (b) suggest potential for corporate data reuse beyond the context of original use. Develop ideal terms and conditions in light of assessment and share broadly.[22]

Distributed Data Science Fluency

Responsible operations depend on broadly distributed data science fluency. Without equal distribution of fluency, libraries lessen the number of organizational inputs into responsible operations. Opportunities for multiple aspects of an organization to contribute to forward progress are many. For example, libraries seeking to use machine learning to improve discovery systems would benefit from teaching and learning librarians and subject liaisons who have been provided with the means (e.g., release time, money, education, and/or experiential opportunities) to gain familiarity with methods and their implications. With foundations like these in place, it becomes possible for projects to be critically championed—or challenged—by staff throughout an organization. A call for fully distributed data science fluency is akin to contemporary efforts that aim to shift a library from a disciplinary to a "functional model"—it is differentiated insofar as the frame of engagement encompasses all roles in an organization in order to leverage the fullest range of experience possible.[23]

Recommendations:

1. Evaluate models within and outside of libraries that support distributed growth of data science fluency throughout an organization. Diversify evaluation according to models that map to a range of organizational resource assumptions (e.g., staffing, budget, and infrastructure). Share the product of evaluation broadly.
2. Pilot distributed data science fluency models; be guided by outcomes of recommendation 1.

Generous Tools

Responsible operations demand the integration of contextual knowledge. Making ready use of contextual knowledge depends in part on the availability of **generous tools**. It is often the case that emerging technologies, the methods they enact, and the variety of programming languages they make use of present a steep learning curve for nonexperts. Use of the technology tends to get siloed to a role in a particular part of the library, and the potential for leveraging diverse forms of expertise present across an organization are lost. Generous tools are designed and documented in such a way that they make it possible for users of varying skill levels to contribute to the improvement and/or use of algorithmic methods. Per Scott Weingart's recommendation, the library community may benefit from seeking out human computer interaction experts to help design generous tools (e.g., human-in-the-loop systems, exploratory visualization environments, GUI-based [graphical user interface] analytics platforms, semiautomated AI model development).[24] These tools could follow in the spirit of Gen (a novice-oriented programming language developed at the Massachusetts Institute of Technology), Zooniverse, and iNaturalist (platforms that can facilitate crowdsourced classification), and resources like those developed by Matthew Reidsma that help librarians audit the product of library discovery systems.[25]

Recommendations:

1. Form a working group focused on studying data science, machine learning, and AI solutions that are designed to accommodate users with varying degrees of technical and methodological experience. Produce high-level synthesis and best practices for solution design in the context of library community need.

2. Foster partnerships between the library developer community, human computer interaction experts, and computer scientists in order to develop systems that are more readily usable by a broad range of library staff.

DESCRIPTION AND DISCOVERY

Digitized and born-digital collection volume and variety present a perennial challenge to library efforts to make collections accessible in a manner that is meaningful to users. Getting from acquisition to access requires significant investment, and resources are often elusive, pushing organizations to seek external funding that fosters labor precarity and uncertain access to collections. Even where resources are abundant, years of collection accumulation and variable content types resist progress. While data science, machine learning, and AI cannot solve these underlying structural problems, they show the potential to create efficiencies that smooth the path to access, enhancing description and expanding forms of discovery along the way.

Enhancing Description at Scale

Discussions of scaling description using computational methods often focus on speed. A less common point of emphasis is the potential for enhancement. Examples are diverse: semantic metadata can be generated from video materials using computer vision; text material description can be enhanced via genre determination or full-text summarization using machine learning; audio material description can be enhanced using speech-to-text transcription; and previously unseen links can be created between research data assets that hold the potential to support unanticipated research questions.[26]

Recommendations:

1. Form a working group focused on assessing algorithmic methods that can be used to enhance description for a range of collection content types. Evaluate open-source and commercial solutions and document extant workflows and staffing requirements that map to a range of organizational realities.
2. Initiate pilots and usability studies informed by outcomes of recommendation 1.

Incorporating Uncertain Description

Attempts to use algorithmic methods to describe collections must embrace the reality that, like human descriptions of collections, machine descriptions come with varying measures of certainty. This should come as no surprise given that algorithms are the product of explicit and latent biases held by humans.

Attempts to use algorithmic methods to describe collections must embrace the reality that, like human descriptions of collections, machine descriptions come with varying measures of certainty.

Challenges in this space are threefold: (1) staff ability to set certainty thresholds, (2) staff ability to incorporate probabilistic data into existing systems without significant modification, and (3) staff ability to explain and contextualize algorithmic description of collections in a manner that is intelligible to the communities they serve.[27] On the last point, the UK National Archives conducted usability testing that is focused on making probabilistic links between records reliably intelligible to a general audience.[28]

Recommendations:

1. Convene a working group and hold multiple symposia focused on probabilistic description, challenges to incorporating probabilistic data into existing systems, and approaches to contextualizing the product of algorithmic methods in a manner that is intelligible to specific user communities. Publish outcomes of working group and symposia openly.
2. Initiate pilots and usability studies informed by outcomes of recommendation 1.

Ensuring Discovery and Assessing Impact

As digital collections and research data grow, libraries face two connected challenges: supporting discovery of collection content on the open web and assessing impact. On the discovery side, institutional repository (IR) managers have sought to optimize their metadata for major search engines. Kenning Arlitsch suggests that machine learning might help the library community train IR content description to an ideal standard, with the ideal standard being used to semiautomate IR content description and remediation.[29] On the impact side, libraries and their peers in museums have released much of their content into the public domain but face challenges assessing the impact of their work. Josh Hadro suggests that cultural heritage organizations might experiment with computer vision in order to identify content reuse on the internet.[30]

Recommendations:

1. Form a working group to investigate computational approaches to assessing content reuse on the open web—working group participation should span galleries, libraries, archives, and museums. The working group is encouraged to consider existing and potential measures of impact.
2. Conduct a study that explores whether machine learning can be used to improve discovery of cultural heritage collections and data assets on the open web.

SHARED METHODS AND DATA

The University of Nebraska–Lincoln applies computer vision to historic newspapers in order to enhance discovery; Indiana University applies natural language processing and machine learning to A/V collections in order to increase access; and the University of North Carolina at Chapel Hill has begun using machine learning to semiautomate systematic reviews in its medical libraries.[31] Collectively, application of algorithmic methods to collections shows promise, yet it is unevenly distributed, and venues, publication outlets, and funding sources for empirically advancing it are rare. In order to broaden the field of participation and improve work in this space, a shift must be made toward shared rather than archipelagic development of algorithmic methods and the data that drives their improvement.

Shared Development and Distribution of Methods

The library community requires interprofessional and interdisciplinary venues, publication outlets, and funding sources that facilitate shared development, implementation, and assessment of algorithmic methods in the context of cultural heritage challenges. A dearth of these resources impacts uptake and refinement. Precedent for venues that inspire future work can be seen in efforts like the Music Information Retrieval Evaluation eXchange (MIREX). MIREX has met since 2005 and focuses meetings around shared tasks. Tasks entail calls for shared methodological refinement in areas like audio fingerprinting, mood and genre classification, and lyrics-to-audio alignment.[32]

Recommendations:

1. Develop venues, publication outlets, and funding sources that facilitate the sharing of methods and benchmarks for machine learning and artificial intelligence.
2. Prototype platforms that facilitate methods competitions specific to cultural heritage contexts (e.g., increased accuracy for handwritten text recognition [HTR]).[33]
3. Launch methods interest groups within and at the junctures of professional and disciplinary associations and societies.

Shared Development and Distribution of Training Data

The viability of machine learning and artificial intelligence is predicated on the representativeness and quality of the data that they are trained on. Organizations with sufficiently representative collections are in a prime position for experimentation. Organizations with less representative collections can have difficulty getting started. Widening the circle of organizational participation 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).[34] Given the tarnished reputation of assumed gold standard training data like ImageNet, this could be a vital contribution to machine learning research. Often presented as a large set of correctly labeled images, ImageNet is predicated upon the biases of many thousands of human contributors—a fact highlighted forcefully by the provocative ImageNet Roulette.[35]

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

Recommendations:

1. Pilot collaborations between institutions with representative collections. Work to share source data and produce "gold standard" training data.
2. Pilot collaborations between institutions with less representative collections. Work to join and share similar source data and produce "gold standard" training data.
3. Conduct a landscape study that focuses on features and requirements of human-in-the-loop (HITL) systems that improve the accuracy of machine learning models (i.e., systems that facilitate a combination of supervised machine learning—curated datasets used to train machine learning algorithms—and active learning—data iteratively tuned by humans to increase accuracy).
4. In addition to Zooniverse, explore whether there are institutions or organizations with sufficiently large user communities that are interested in implementing and managing a platform that facilitates the creation of corpora and classification of collections from smaller, less well-resourced organizations.[36]

MACHINE-ACTIONABLE COLLECTIONS

Machine-actionable collections (alternatively, collections as data) 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 for Indigenous Data Governance").[37] Users of these collections span library staff and the communities they serve. For example, on the library side, computational work to enhance discovery is predicated on the ready availability of machine-actionable collections. On the researcher side, projects

that conduct analysis at scale are similarly dependent on ready access to machine-actionable collections. Development of machine-actionable collections should be guided by clearly articulated principles and ethical commitments—"The Santa Barbara Statement on Collections as Data" provides one resource to work through considerations in this space.[38] Overall, three high-level challenges characterize research agenda contributor input on this topic: (1) making machine-actionable collections a core activity; (2) broadening machine-actionable collections; and (3) rights assessment at scale.

Making Machine-Actionable Collections a Core Activity

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. 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.[39] Furthermore, workflows anchored in core activity will begin to show the potential of algorithmic methods to assist with processing collections at scale; alleviate concerns about sustainability by proving impact on core operations; and help smooth the path to integrating probabilistic data in a discovery system—a challenge that vexes many libraries.[40]

Recommendations:

1. Initiate coordinated user studies, at an international level, that work toward standardizing multiple levels of machine-actionable collection need. Studies are guided by user experience and human-computer interaction principles.[41]
2. Use the product of recommendation 1 to develop requirements for base-level machine-actionable collections.
3. Develop workflows that leverage data science, machine learning, and AI to help process digital collections at scale (e.g., scene segmentation, objects in images, speech-to-text transcription).
4. Develop workflows in partnership with disciplinary researchers that can identify, extract, and make machine-actionable data from general and special collections to fuel library experimentation and research activity on campus (e.g., handwriting to full text, hand-drawn tables of numeric data to structured data, herbarium specimens).[42]
5. Identify opportunities for data "loops" (e.g., the product of a crowdsourcing platform is used to enhance general discovery and provide training data to fuel machine learning).[43]

Historic and contemporary biases in collection development activity manifest as corpora that overrepresent dominant communities and underrepresent marginalized communities.

Broadening Machine-Actionable Collections

While the number of collections tuned for computation grows, it remains the case that the majority are the product of large Western institutions. Historic and contemporary biases in collection development activity manifest as corpora that overrepresent dominant communities and underrepresent marginalized communities. Where marginalized communities are

represented, that representation tends to be within the context of narratives that dominant cultures sanction.[44] A critical historical perspective and resources are required to create corpora that remediate underrepresentation. Without these steps, libraries and researchers run the risk of reifying existing biases in a limited cultural record.

Beyond the question of limited community representation, machine-actionable collections also tend to be text data expressed predominantly in English. A lack of linguistic diversity in machine-actionable collections limits library and research community potential. Fields like natural language processing are severely constrained by this reality, a state of play that requires self-regulation via application of the "Bender Rule."[45] Broadening content type availability beyond text to images, moving images, audio collections, web archives, social media data, and things like scientific special collections (e.g., 18th-century weather observations, specimens) would foster greater library and research community possibilities. Broadened content type availability calls for the development of policies, practices, and platforms that navigate rights and terms and conditions associated with these collections. With respect to potential solutions, Taylor Arnold and Lauren Tilton have suggested that having an approach like JSTOR's Data for Research (DFR) for A/V content would be helpful, and Ed Summers has similarly suggested that something like the HathiTrust Research Center's data capsule could help facilitate social media data collection use.

Recommendations:

1. Prioritize 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.[46]
2. Prioritize the creation of machine-actionable collections with greater linguistic diversity.
3. Convene working groups and pilots to explore policy and infrastructure requirements for providing access to and/or supporting analysis of machine-actionable collections that are inclusive of less available content types (e.g., audio, video, social media)—draw inspiration from efforts like JSTOR's DFR and the HathiTrust Research Center's data capsule and extend to efforts like Documenting the Now, Project AMP, and the Distant Viewing Lab.[47]

Rights Assessment at Scale

Rights assessment at scale presents significant challenges for libraries. The prospect of machine-actionable collection use compounds difficulties: users 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 (e.g., openly licensed Flickr photos of minors re-used years later, without consent, to improve surveillance technology).[48] Collectively, these challenges present a "wicked problem" for the library community.[49] Building on past work, and engaging with contemporary efforts like the New York Public Library's *Unlocking the Record of American Creativity: The Catalog of Copyright Entries, Building Legal Literacies for Text Data Mining (Building LLTDM)*, Bergis Jules' work on consent and social media data use, and the Global Indigenous Data Alliance's "CARE Principles for Indigenous Data Governance" will help the library community develop a range of strategies to help address these challenges.[50]

Recommendations:

1. Form a working group that investigates current and potential strategies for addressing rights assessment at scale. In combination, this work should investigate current and potential strategies for ensuring the ethical use of collections. This combination is essential—legal use does not equal ethical use.

WORKFORCE DEVELOPMENT

A tool has no impact without a hand to guide it. The same logic extends to data science, machine learning, and AI. The library community works to give these technologies and methods purpose in alignment with their values. Some within the space already do, but the capacity to do so is unevenly distributed. In order to address this imbalance, a range of workforce development challenges lie ahead. High-level challenges identified by contributors to this agenda include investigating core competencies, committing to internal talent, and evidence-based training.

Investigating Core Competencies

Workforce development geared toward data science, machine learning, and AI capacity building requires determining what combination of competencies, experiences, and dispositions will support the directions that libraries are seeking to take.[51] On the subject of dispositions, agenda contributors suggest that the ability to translate domain knowledge and technical knowledge between communities with varying degrees of expertise will be crucial. Given that critical use of these technologies and methods requires experiences that accrue to a broad range of expertise, some agenda contributors suggest removal of library science degree requirements for library staff and faculty positions. Candidates with these skills will likely be in demand across sectors and it may be the case that libraries cannot compete on salary. In lieu of competition on salary, libraries should investigate other means of competition (e.g., remote work as a normative option).[52] 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. All of the above raises the question of administrative competencies that effectively guide, integrate, and sustain data science, machine learning, and AI work in libraries.

Recommendations:

1. Investigate core competencies, experiences, and dispositions that the library community believes are essential to data science, machine learning, and/or AI efforts in libraries. Investigation should span development of requirements for library staff *and* the administrators responsible for guiding, integrating, and sustaining this work.
2. Use the product of recommendation 1 to inform curricular development in graduate programs and ongoing professional development opportunities for library staff and administrators.

Committing to Internal Talent

Emerging technology and innovation tend to be the province of staff brought in from outside an organization. This begs the question of why it seems to be the case that it is less common to support reshaping existing roles and responsibilities. The answer may be that it is easier to hire someone new, but contributors to the agenda expressed strong desire for commitment 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.

Recommendations:

1. Form a working group to investigate the development of organizational models that avoid silos and support hybridity between core and emerging services; models of this kind may encourage natural diversification and/or deepening of skills over time.

Expanding Evidence-based Training

Data science, machine learning, and AI training options are many, yet as Kate Zwaard notes, few are evidence based.[53] It can be difficult to assess which training options are most likely to develop desired skills. *The Carpentries* present a notable exception that can inspire additional efforts in this space.[54] Beyond being evidence based, research agenda contributors stressed that training options should be low cost or free. While this might ensure more equal opportunity, it raises the challenge of sustainability, quality, and fair compensation. Finally, contributors also expressed strong desire for training options grounded in library use cases—the British Library, National Archives UK, and Birkbeck University's *Computing for Cultural Heritage* suggest promising efforts in this vein.[55]

Recommendations:

1. Initiate evidence-based evaluations of existing data science, machine learning, and AI training opportunities within and outside of the library community.
2. Pilot and/or support the development of evidence-based data science, machine learning, and/or AI training options that are grounded in library use cases.
3. Explore, document, develop, and share sustainability models for keeping training opportunities free or low cost without sacrificing quality and fair compensation.

DATA SCIENCE SERVICES

Modeling Data Science Services

To support the viability of data science services at a range of institutions, the profession is in need of service plans that can guide libraries with different resources, staffing, and missions.[56] Without diversification there is a risk of pricing big swaths of the library community out of the conversation.[57] Beyond single institution solutions, the library community might also take inspiration from efforts like the *Data Curation Network* and *Project CADRE*—multi-institutional models that load-balance social and technical dimensions of library services.[58]

Without diversification there is a risk of pricing big swaths of the library community out of the conversation.

Aspirations to scale and service complexity should be grounded by nuanced evaluation of user needs. Harriett Green has suggested that the profession may benefit from studies that investigate "the long tail" of data science support (e.g., the portion of minimally resource-intensive data science support requests, that when added up, result in an equal or even greater resource demand than exceptional, highly resource-intensive requests).[59] Studies of this kind could help inform data science resource assumptions, charting paths that are feasible for a wider range of institutions.

Recommendations:

1. Initiate a call for data science service case studies that describe the current state of practice.

2. Initiate a call for "long tail" of data science support studies grounded in different institutional contexts.
3. Initiate a working group focused on formalizing a set of data science service plans. Plans should be developed in such a way that a range of institutions can use them.
4. Initiate a multi-institutional effort to develop data science personas that span library staff, the communities they collaborate with, and the communities they serve.

Research and Pedagogy Integration

Integrating data science, machine learning, and AI with library research support and pedagogy presents a number of opportunities. In some cases, these technologies and methods create efficiencies that free up time for library staff to more deeply integrate with a broader disciplinary research ecosystem.[60] Consider the example of *The Space Library*—a collaboration between Wolbach Library, Libre Space Foundation, and NASA's Small Spacecraft Systems Virtual Institute focused on, "[developing] open metadata standards and enabling public engagement with space technology in public libraries."[61] In other cases, potential afforded by technologies and methods is about enhancing the value of an existing service or presenting an opportunity to fill a gap. Consider the University of North Carolina at Chapel Hill Libraries' use of machine learning to develop more comprehensive systematic reviews and Columbia University Libraries' work to centralize and scale research computing services at the basic level—a significant contribution to a multilayered conception of campus-wide research support.[62]

Libraries could also consider using information literacy instruction as a vector to introduce algorithmic concepts and their ethical implications—Blake Payne's "AI + Ethics Curriculum for Middle School" presents a model that could be adapted to a library context.[63] That kind of effort aligns with Sofia Leung, Michelle Baildon, and Nicholas Albaugh's argument that reference, instruction, and collections work can be used to promote algorithmic justice (e.g., exploring critical development of AI-assisted reference systems, reimagining the ways that instruction is assessed, and exploring partnerships with marginalized communities that support the potential creation of more representative training data).[64]

Recommendations:

1. Initiate a working group focused on documenting opportunities afforded by data science, machine learning, and AI to more deeply integrate library staff within the campus research ecosystem.
2. Per Sofia Leung, Michelle Baildon, and Nicholas Albaugh, initiate multiple efforts that apply algorithmic justice in reference, instruction, and collections work.

Libraries need people—appropriately resourced—who can translate, align, and balance a wide range of interests.

SUSTAINING INTERPROFESSIONAL AND INTERDISCIPLINARY COLLABORATION

The work that lies ahead stands to benefit from well-conceived collaborations within and across professional and disciplinary communities.Individuals within each have expertise born of particular contexts, and context is key to critical use of data science, machine learning, and AI. Forging and sustaining these collaborations can be challenging. Libraries need people—appropriately resourced—who can translate, align, and balance a wide range of interests. In some cases, as Lauren Tilton notes, interest may be readily convergent between general library practice and the aims of a discipline-like public history. Nonetheless, the sustainability of ongoing collaborations can be troubled as they crash into disciplinary funding silos.

For example, if a cultural heritage practitioner and a computer scientist seek funding from a science-focused funding source, they may be told that enhancements to archival discovery do not constitute an advance of basic research. On the other side, applications to a humanities-focused funding source may be told that their project lacks a humanities research question. The funding challenge for cultural heritage organizations is not intractable—National Science Foundation support was fundamental to the development of digital libraries.[65] Whatever shape professional and disciplinary collaborations take they should be resourced in a manner that supports all parties in the collaboration.

Recommendations:

1. Form a practitioner and funder working group—broadly conceived across public and private funders—to assess cultural heritage funding challenges encountered at the intersection between professional and disciplinary research questions.
2. Investigate the potential for collective investment in applied technical research akin to David Lewis' 2.5% commitment. Consider adopting a sliding scale to encourage inclusive organizational participation. Fund cross-functional community collaborations that advance data science, machine learning, and AI in a library context.[66]
3. Develop cross-functional community groups that are anchored in and/or sit at the juncture of multiple associations and societies. Orient collaboration around specific initiatives (e.g., ethical AI, training data development, methodological development, domain concept primers).
4. Support cross-functional community research and practice sprints that seek to apply data science, machine learning, and AI to library challenges.

NEXT STEPS

Responsible Operations presents an interdependent set of technical, organizational, and social challenges to be addressed en route to library operationalization of data science, machine learning, and artificial intelligence. It is our hope that *Responsible Operations* resonates with, amplifies, and helps make a case for the work that lies before us. Per the introduction, no single country, association, or organization can meet these challenges.

We all have a role to play.

With investigations that span collections and research support, the scholarly record, the system-wide library, user studies, community catalysts, and data science, OCLC Research will assess how it might best contribute. This assessment will take place in light of our capacity, existing research library community efforts, and expressed community needs. We thank our advisory board and our landscape group for their generous contributions to this work. Without them, this initial research agenda would not have been possible.

ACKNOWLEDGEMENTS

The advisory group helped lay the foundation for the agenda, introduced challenges and opportunities, and provided feedback. The author is grateful for their guidance.

- Kenning Arlitsch, Montana State University
- Jon Cawthorne, Wayne State University
- Karen Estlund, Colorado State University
- Josh Hadro, IIIF Consortium
- Bohyun Kim, University of Rhode Island
- Trevor Owens, Library of Congress
- Benjamin Schmidt, New York University
- Sarah Shreeves, University of Arizona
- MacKenzie Smith, University of California, Davis
- Claire Stewart, University of Nebraska – Lincoln
- Melissa Terras, University of Edinburgh
- Diane Vizine-Goetz, OCLC
- John Wilkin, University of Illinois at Urbana-Champaign
- Kate Zwaard, Library of Congress

The landscape group provided the primary contributions used to form the agenda. Contributions came via interviews, participation in a two-day event in Dublin, OH—"Shaping an Applied Research Agenda"—and asynchronous feedback. The agenda would not be possible without their generosity.

- Ruth Ahnert, Queen Mary University of London
- Taylor Arnold, University of Richmond
- Helen Bailey, Massachusetts Institute of Technology
- Ted Baldwin, University of Cincinnati
- Daina Bouquin, Harvard University and the Smithsonian Institution
- Karen Cariani, WGBH
- Michelle Cawley, University of North Carolina at Chapel Hill
- Rumman Chowdhury, Accenture
- Jason Clark, Montana State University
- Nicole Coleman, Stanford University
- Rebecca Dikow, Smithsonian Institution
- Quinn Dombrowski, Stanford University
- Virginia Dressler, Kent State University
- Jon Dunn, Indiana University
- Ixchel Faniel, OCLC

- Maggie Farrell, University of Nevada, Las Vegas
- Lisa Federer, National Institutes of Health
- Barbara Fister, Gustavus Adolphus College
- Kathleen Fitzpatrick, Michigan State University
- Themba Flowers, Yale University
- Alex Gil, Columbia University
- Jean Godby, OCLC
- Tiffany Grant, University of Cincinnati
- Jane Greenberg, Drexel University
- Harriett Green, Washington University in St. Louis
- Umi Hsu, ONE Archives Foundation
- Richard Johansen, University of Cincinnati
- Bohyun Kim, University of Rhode Island
- Lauren Klein, Georgia Tech
- Emily Lapworth, University of Nevada, Las Vegas
- Shari Laster, Arizona State University
- Matthew Lincoln, Carnegie Mellon University
- Meris Longmeier, The Ohio State University
- Dominique Luster, Carnegie Museum of Art
- Nandita Mani, University of North Carolina at Chapel Hill
- Sara Mannheimer, Montana State University
- Richard Marciano, University of Maryland
- Alexandra Dolan-Mescal, Harvard University
- David Minor, University of California, San Diego
- Marilyn Myers, University of Houston
- Peace Ossom Williamson, University of Texas at Arlington
- Carole Palmer, University of Washington
- Merrilee Proffitt, OCLC
- Christopher Prom, University of Illinois at Urbana-Champaign
- Matthew Reidsma, Grand Valley State University
- Mia Ridge, British Library
- Danielle Robinson, Code for Science & Society
- Barbara Rockenbach, Columbia University
- Amanda Rust, Northeastern University
- Yasmeen Shorish, James Madison University
- David Smith, Northeastern University
- Ed Summers, University of Maryland

- Santi Thompson, University of Houston
- Jer Thorp, Library of Congress
- Lauren Tilton, University of Richmond
- Ted Underwood, University of Illinois at Urbana-Champaign
- Chela Scott Weber, OCLC
- Keith Webster, Carnegie Mellon University
- Scott Weingart, Carnegie Mellon University
- Jon Wheeler, University of New Mexico
- Stanley Wilder, Louisiana State University
- Jamie Wittenberg, Indiana University
- Scott Young, Montana State University

NOTES

1. Brown, Adrienne M. 2017. *Emergent Strategy.* Chico, CA: AK Press.

2. This work focuses on the research library community and runs alongside contemporary efforts like: Kennedy, Mary Lee. 2019. "What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?" *Research Library Issues*, no. 299 (September): 3–13. https://doi.org/10.29242/rli.299.1;

   Cox, Andrew M., Stephen Pinfield, and Sophie Rutter. 2018. "The Intelligent Library: Thought Leaders' Views on the Likely Impact of Artificial Intelligence on Academic Libraries." *Library Hi Tech* 37 No. 3: 418-435. https://doi.org/10.1108/LHT-08-2018-0105.

3. The author thanks Rumman Chowdhury for introducing him to the concept of responsible operations. The author subsequently adapted and expanded the concept in the context of library community need. For more on Rumman Chowdhury's work see http://www.rummanchowdhury.com/.

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

   Whittaker, Meredith, and Kate Crawford. 2019. "AI in 2019: A Year in Review." *Medium*. 9 October 2019. https://medium.com/@AINowInstitute/ai-in-2019-a-year-in-review-c1eba5107127.

5. The advisory group is comprised of senior library leaders and disciplinary scholars. The landscape group is predominantly comprised of library staff. Additional representation in the landscape group is drawn from museum, archive, open science, disciplinary, and professional consulting communities;

   Padilla, Thomas. "Shaping an Applied Research Agenda." *Hanging Together* (blog), OCLC Research, 17 May 2019. https://hangingtogether.org/?p=7320.

6. Chowdhury defines responsible operations as collective investments in, " . . . processes to combat algorithmic bias." See Apte, Poornima. 2017. "The Data Scientist Putting Ethics Into AI." *The Daily Dose, OZY*, 25 September 2017. http://www.ozy.com/rising-stars/rumman-chowdhury-the-human-centric-thinker/81044.

7. Adler, Melissa. 2017. *Cruising the Library: Perversities in the Organization of Knowledge*. First edition. New York: Fordham University Press;

   Anderson, Jane, and Kimberly Christen. 2019. "Decolonizing Attribution." *Journal of Radical Librarianship* 5 (June): 113–52;

   Jones, Michael. 2019. "Collections in the Expanded Field: Relationality and the Provenance of Artefacts and Archives." *Heritage* 2 (1): 884–97. https://doi.org/10.3390/heritage2010059.

8. Clark, Jason A. (2018) 2019. "Home for the IMLS Grant RE-72-17-0103-17 - 'RE:Search' - Unpacking the Algorithms That Shape Our UX." jasonclark/algorithmic-awareness, GitHub. Accessed 20 November 2019. https://github.com/jasonclark/algorithmic-awareness;

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

9. Chowdhury, Rumman, Lade Obamehinti, and Rodney Sampson (Moderator). 2019. "Diversity, Equity, and Inclusion is Imperative in AI Design." Panel discussion during Diversity, Equity, and Inclusioin, Business AI Integration, VB Transform 2019 Conference. Produced by VentureBeat. Posted 15 July 2019. YouTube video 25:59 https://www.youtube.com/watch?v=xmtBMes0uYY.

10. Floridi, Luciano, and Josh Cowls. 2019. "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review,* 1 (1). https://doi.org/10.1162/99608f92.8cd550d1.

11. Hicks, Marie. 2019. "Hacking the Cis-Tem." *IEEE Annals of the History of Computing* 41 (1): 20–33. https://doi.org/10.1109/MAHC.2019.2897667;

    Ewing, Tom. n.d. "Data in Social Context." Accessed 26 August 2019. https://sites.google.com/vt.edu/etewing/disc.

12. Sherratt, Tim. (2012) 2019. "The Real Face of White Australia: Experimental Browser. National Archives of Australia. Accessed 13 August 2019. http://invisibleaustralians.org/faces/.

13. Buolamwini, Joy. 2016 "How I'm Fighting Bias in Algorithms." Filmed November 2016 in Boston, Massachusetts. TEDxBeaconStreet video, 08:37. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms;

    Gershgorn, Dave. 2019. "A Privacy Dustup at Microsoft Exposes Major Problems for A.I.: The most important Dataset are Rife with Sexism and Racism" *OneZero* (blog), 12 June 2019. https://onezero.medium.com/a-privacy-dustup-at-microsoft-exposes-major-problems-for-ai-53e0b4206e98;

    Snow, Jacob. 2018. "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots." *American Civil Liberties Union* (blog), 26 July 2018. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

14. Nicole Coleman, Interview, 1 April 2019.

15. Peet, Lisa. 2016. "Library of Congress Drops Illegal Alien Subject Heading, Provokes Backlash Legislation." *Library Journal* (news), 13 June 2016. https://www.libraryjournal.com?detailStory=library-of-congress-drops-illegal-alien-subject-heading-provokes-backlash-legislation.

16. West, Sarah Myers, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race and Power in AI*. New York: AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf.

17. Metcalf, Jacob, Emanuel Moss, and danah boyd. 2019. *Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics*. New York: The New School. https://datasociety.net/wp-content/uploads/2019/09/Owning-Ethics-PDF-version-2.pdf. [First appeared in *Social Research: An International Quarterly* 82, no. 2, (Summer): 449-476. https://muse.jhu.edu/article/732185.]

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

19. IBM Research Trusted AI. n.d. "AI Explainability 360 Open Source Toolkit." Accessed 12 August 2019. aix360.mybluemix.net;

Caruana, Rich, Harsha Nori, Samuel Jenkins, Paul Koch, and Ester de Nicolas. 2019. "Creating AI Glass Boxes – Open Sourcing a Library to Enable Intelligibility in Machine Learning." *Microsoft Research Blog*, 10 May 2019. https://www.microsoft.com/en-us/research/blog/creating-ai-glass-boxes-open-sourcing-a-library-to-enable-intelligibility-in-machine-earning/.

20. Mia Ridge noted that the Bodleian Libraries accept catalog pull requests via Github. See Bodleian/Medieval-Mss. (2017) 2019. HTML. Bodleian Libraries. GitHub. https://github.com/bodleian/medieval-mss; Matthew Lincoln noted that practices of this kind could increase workload for staff and should be planned for accordingly.

21. Jer Thorp, interview, 22 March 2019.

22. Amanda Rust, interview, 29 August 2019.

23. Burton, Matt, Liz Lyon, Chris Erdmann, and Bonnie Tijerina. 2018. "Shifting to Data Savvy: The Future of Data Science In Libraries." Monograph (Project Report). University of Pittsburgh, Data & Society, and North Carolina State University. http://d-scholarship.pitt.edu/33891/;

Frenkel, Ann, Tiffany Moxham, Dani Brecher Cook, and Brianna Marshall. 2018. "Moving from Subject Specialists to a Functional Model." *Research Library Issues*, no. 294: 39–71. https://doi.org/10.29242/rli.294.5;

Hickerson, Thomas and John Brosz. 2019. "Remaining Relevant: Critical Roles for Libraries in the Research Enterprise." Paper presented at the 85th World Library and Information Congress of IFLA, Athens, Greece, 25 August 2019. http://library.ifla.org/2575/1/082-hickerson-en.pdf.

24. D'Onfro, Jillian. 2019. "DataRobot Becomes a Unicorn by Selling AI Toolkits to Harried Data Scientists." *Forbes*, 17 September 2019. https://www.forbes.com/sites/jilliandonfro/2019/09/17/machine-learning-startup-datarobot-raises-206-million/.

25. Cusumano-Towner, Marco F., Feras A. Saad, Alexander K. Lew, and Vikash K. Mansinghka. 2019. "Gen: A General-Purpose Probabilistic Programming System with Programmable Inference." In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* (PLDI), Phoenix, AZ, 22-26 June 2019, 221–236. New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3314221.3314642;

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

Koch, Julian, and Simon Stisen. 2017. "Citizen Science: A New Perspective to Advance Spatial Pattern Evaluation in Hydrology." *PLOS ONE* 12 (5): e0178165. https://doi.org/10.1371/journal.pone.0178165;

O'Brien, Colleen. 2017. "App Combines Computer Vision and Crowdsourcing to Explore Earth's Biodiversity, One Photo at a Time." *Mongabay Environmental News* 30 August 2017. https://news.mongabay.com/2017/08/smartphone-app-combines-computer-vision-and-crowdsourcing-to-explore-earths-biodiversity-one-photo-at-a-time/;

Reidsma, Matthew. 2019. *Masked by Trust: Bias in Library Discovery.* Sacramento, CA: Library Juice Press.

26. Karen Cariani, interview, 24 June 2019;

    Ted Underwood, interview, 11 March 2019;

    Keith Webster, interview, 25 March 2019;

    Distant Viewing Lab. n.d. "Analyzing Visual Culture." Accessed 21 August 2019. https://distantviewing.org;

    Dunn, Jon W., Juliet L. Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf. 2018. *Audiovisual Metadata Platform (AMP) Planning Project: Progress Report and Next Steps*. AVP, Indiana University, and the University of Texas at Austin. https://scholarworks.iu.edu/dspace/handle/2022/21982;

    Hejblum, Boris P., Griffin M. Weber, Katherine P. Liao, Nathan P. Palmer, Susanne Churchill, Nancy A. Shadick, Peter Szolovits, Shawn N. Murphy, Isaac S. Kohane, and Tianxi Cai. 2019. "Probabilistic Record Linkage of De-Identified Research Datasets with Discrepancies Using Diagnosis Codes." *Scientific Data* 6 (January): 180298. https://doi.org/10.1038/sdata.2018.298;

    Taylor, Arnold, and Lauren Tilton. 2019. "Distant Viewing: Analyzing Large Visual Corpora." *Digital Scholarship in the Humanities* 0 (0). https://doi.org/10.1093/digitalsh/fqz013.

27. Helen Bailey, interview, 2 July 2019;

    Benjamin Schmidt, interview, 19 March 2019.

28. Ranade, Sonia. 2016. "Traces through Time: A Probabilistic Approach to Connected Archival Data." In *2016 IEEE International Conference on Big Data (Big Data)*, 3260–65. Washington DC: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData.2016.7840983.

29. Kenning Arlitsch, interview, 7 March 2019.

30. Josh Hadro, interview, 12 March 2019;

    Machine Learning forArtists (ml4a). (2016) 2019. "Practical Guides, Tutorials, and Code Samples for ml4a." ml4a/ml4a-guides. (Language: Jupyter Notebook). GitHub. https://github.com/ml4a/ml4a-guides.

31. Michelle Cawley, interview, 20 August 2019;

    Dunn et al., "Audiovisual Metadata Platform (AMP)," (see note 26);

    Lorang, Elizabeth, Leen-Kiat Soh, Chulwoo Pack, Yi Liu, Delaram Rahimighazikalayeh, and John O'Brien. 2019. "Application of the Image Analysis for Archival Discovery Team's First-Generation Methods and Software to the Burney Collection of British Newspapers." CDRH Grant Reports, May. https://digitalcommons.unl.edu/cdrhgrants/7.

32. Mirex: https://www.music-ir.org/mirex/wiki/MIREX_HOME#Welcome_to_MIREX_2019.

33. Kaggle: https://www.kaggle.com/competitions.

34. Rumman Chowdhury, comment on a draft version of Thomas Padilla's 2019 "Shaping an Applied Research Agenda." *Hanging Together* (blog), *OCLC Research,* 17 May 2019. https://hangingtogether.org/?p=7320.

35. Boykis, Vicki. 2019. "Neural Nets Are Just People All the Way Down: Is Machine Learning Really Powered by Machines? (No)." *Normcore Tech,* 16 October 2019. https://vicki.substack.com/p/neural-nets-are-just-people-all-the;

    Crawford, Kate, and Trevor Paglen. 2019. "Excavating AI: The Politics of Images in Machine Learning Training Sets." *AI Now Institute, NYU* 19 September 2019. https://www.excavating.ai;

    Metz, Cade. 2019. "'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?" *New York Times* (Art & Design), 20 September 2019. https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html.

36. LibraryOfCongress/Concordia. (2018) 2019. "Crowdsourcing Platform for Full Text Transcription and Tagging." (Python). GitHub. https://github.com/LibraryOfCongress/concordia;

    Zooniverse. "Welcome to the Zooniverse." https://www.zooniverse.org/.

37. Padilla, Thomas. 2017. "On a Collections as Data Imperative." *UC Santa Barbara*. (February). https://escholarship.org/uc/item/9881c8sv;

    Padilla, Thomas. 2016. "Humanities Data in the Library: Integrity, Form, Access." *D-Lib Magazine* 22, no. 3/4 (March). https://doi.org/10.1045/march2016-padilla;

    Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. *Always Already Computational: Collections as Data—Final Report*. Zenodo. https://doi.org/10.5281/zenodo.3152935;

    Research Data Alliance International Indigenous Data Sovereignty Interest Group. 2019. "CARE Principles for Indigenous Data Governance." Global Indigenous Data Alliance. https://www.gida-global.org/care;

    "Traditional Knowledge (TK) Labels." 2019. Local Contexts. (Archived 7 Feb 2016 - 15 Nov 2019) https://web.archive.org/web/20191009065200/https:/localcontexts.org/tk-labels/.

38. Padilla, *Collections as Data* (see note 37).

39. Trevor Owens, interview, 5 March 2019.

40. Mia Ridge, interview, 22 March 2019.

41. Trevor Owens, interview, 27 March 2019.

42. MacKenzie Smith, interview, 7 March 2019

    "Kuzushiji Recognition: Opening the Door to a Thousand Years of Japanese Culture." 2019. kaggle. https://kaggle.com/c/kuzushiji-recognition;

Schuettpelz, Eric, Paul Frandsen, Rebecca Dikow, and Laurence Dorr. 2017. "Applications of Deep Convolutional Neural Networks to Digitized Natural History Collections." *Biodiversity Data Journal* 5 (February): e21139. https://doi.org/10.3897/BDJ.5.e21139;

Transkribus. n.d. "Transcribe. Collaborate. Share...: ...and Benefit From Cutting Edge Research in Handwritten Text Recognition!" Accessed 20 November 2019. https://transkribus.eu/Transkribus/;

Whitmire, Amanda. 2019. "CalCOFI Hydrobiological Survey of Monterey Bay." Always Already Computational - Collections as Data. GitHub. https://collectionsasdata.github.io/facet3/.

43. Trevor Owens, interview, 27 March 2019;

    Mia Ridge notes this is a goal of the British Library's 2019 "Living with Machines: A Research Project Combining Digital Archives With Data Science Techniques to Investigate the Effects of Mechanisation on Society." 13 November 2019. https://www.bl.uk/projects/living-with-machines.

44. This can be forcefully expressed in digitization priorities. See: Ziegler, S. L. 2019. "Digitization Selection Criteria as Anti-Racist Action." *Code4Lib Journal*, no. 45 (August 9, 2019). https://journal.code4lib.org/articles/14667.

45. Bender, Emily. 2019. "The #BenderRule: On Naming the Languages We Study and Why It Matters." *The Gradient*. 14 September 2019. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

46. Digital Library Federation. 2019. "DLF Cultural Assessment Working Group." https://wiki.diglib.org/Assessment:Cultural_Assessment#DLF_Cultural_Assessment_Working_Group;

    Northeastern University Digital Scholarship Group. n.d. "Design for Diversity: An IMLS National Forum Project." Accessed 20 November 2019. https://dsg.neu.edu/research/design-for-diversity/;

    Digital Justice Lab. n.d. "Call for Participation: Please Don't Include Us. We Are Now Accepting *Participant Applications* for Our Please Don't Include Us Workshop!" Tides Canada. Accessed 20 November 2019. https://digitaljusticelab.ca/cfp.

47. Arnold, Taylor, and Lauren Tilton. "Distant Viewing Lab." https://distantviewing.org/

    DocNow. n.d. "DocNow is a Tool and a Community Developed Around Supporting the Ethical Collection, Use, and Preservation of Social Media Content." Accessed 20 November 2019. https://www.docnow.io/;

    Dunn, Jon W., Juliet L. Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf. 2018. "Audiovisual Metadata Platform (AMP) Planning Project: Progress Report and Next Steps." 28 March 2018. https://scholarworks.iu.edu/dspace/handle/2022/21982.

48. Hill, Kashmir, and Aaron Krolik. "How Photos of Your Kids Are Powering Surveillance Technology." *The New York Times* (Technology Section), 11 October 2019. https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html;

    University of Washington, UW Graphics and Imaging Laboratory. 2015. "MegaFace and MF2: Million-Scale Face Recognition" MegaFace. http://megaface.cs.washington.edu/.

49. Kolko, Jon. 2012. "Wicked Problems: Problems Worth Solving." *Stanford Social Innovation Review*. 6 March 2012. https://ssir.org/articles/entry/wicked_problems_problems_worth_solving.

50. Association of Research Libraries, Center for Social Media, School of Communication, American University, Washington College of Law, American University. 2012. *Code of Best Practices in Fair Use for Academic and Research Libraries*. Washington DC: ARL and American University. https://publications.arl.org/code-fair-use/;

Research Data Alliance International Indigenous Data Sovereignty Interest Group. 2019. "CARE Principles for Indigenous Data Governance." Global Indigenous Data Alliance. https://www.gida-global.org/care;

Cram, Greg. 2018. "Unlocking the Record of American Creativity—with Your Help." *New York Public Library* 30 March 2018. https://www.nypl.org/blog/2018/03/30/unlocking-record-american-creativity;

Jules, Bergis. "Some Thoughts on Ethics and DocNow." DocNow. Medium. 3 June 2016. https://news.docnow.io/some-thoughts-on-ethics-and-docnow-d19cfec427f2;

Samberg, Rachael. "Team Awarded Grant to Help Digital Humanities Scholars Navigate Legal Issues of Text Data Mining." UC Berkeley Library Update. 14 August 2019. https://update.lib.berkeley.edu/2019/08/14/team-awarded-grant-to-help-digital-humanities-scholars-navigate-legal-issues-of-text-data-mining/.

51. Karen Estlund, interview, 13 March 2019.

52. Padilla, Thomas. 2019. "Shaping an Applied Research Agenda." *Hanging Together* (blog), OCLC Research, 17 May 2019. https://hangingtogether.org/?p=7320.

53. Kate Zwaard, interview, March 14, 2019.

54. The Carpentries. Assessment and Impact." The Carpentries. Accessed 20 November 2019. https://carpentries.org/assessment/.

55. McGregor, Nora. 2019. "Computing for Cultural Heritage: A Project to Develop a New Postgraduate Certificate, Computing for Cultural Heritage, with The National Archives and Birkbeck University." The British Library, Projects. https://www.bl.uk/projects/computingculturalheritage.

56. Bohyun Kim, interview, 19 March 2019.

57. Sarah Shreeves, interview, April 1, 2019.

58. John Wilkin, interview, March 12, 2019;

"CADRE: Collaborative Archive & Data Research Environment." n.d. Indiana University Network Science Institute. Accessed 20 November 2019. https://iuni.iu.edu/resources/cadre.

The Data Curation Network. "We are the Data Curation Network." https://datacurationnetwork.org/.

59. Harriett Greene, interview, 22 March 2019.

60. Daina Bouquin, interview, 24 June 2019.

61. Harvard & Smithsonian Center for Astrophysics. n.d. “The Space Library.” Wolback Library. Accessed 20 November 2019. https://library.cfa.harvard.edu/space-library.

62. Michelle Cawley, interview, 20 August 2019;

    Barbara Rockenbach, interview, 28 July 2019.

63. Peace Ossom-Williamson, interview, 13 March 2019;

    Barbara Fister, interview, 11 March 2019;

    Payne, Blakeley H. n.d. “Project: AI + Ethics Curriculum for Middle School.” Overview. MIT Media Lab. Massachusettes Institute of Technology. Accessed 20 November 2019. https://www.media.mit.edu/projects/ai-ethics-for-middle-school/overview/.

64. Leung, Sofia, Michelle Baildon, and Nicholas Albaugh. 2019. “Applying Concepts of Algorithmic Justice to Reference, Instruction, and Collections Work.” *DSpace@MIT* 30 September 2019. https://dspace.mit.edu/handle/1721.1/122343.

65. Griffin, Stephen. 1998. “NSF/DARPA/NASA Digital Libraries Initiative: A Program Manager’s Perspective.” *D-Lib Magazine* (July/August). http://www.dlib.org/dlib/july98/07griffin.html.

66. Lewis, David W. 2017. “The 2.5% Commitment.” ScholarWorks Repository. IUPUI University Library. 11 September 2017. https://doi.org/10.7912/c2jd29.

APPENDIX: TERMS

Basic definitions of data science, machine learning, and artificial intelligence are provided below. The scope of terms included below is not exhaustive and definitions trend toward the general rather than the specific.

Data science refers to a broad set of methods used to extract, analyze, and communicate insights about patterns in data at scale. The impact of data science is felt in many aspects of life. Data science increasingly drives decision-making (e.g., investing), products (e.g., recommender systems—books, music, movies), and services (e.g., healthcare). According to John Kelleher and Brendan Tierney, data science has roots in the history of data collection and data analysis—fields of activity with challenges compounded by the convergence of computation and digitization. Subsequent growth in data generation prompted innovations in data storage (relational data models, structured query language, nonrelational data models) and data analysis (machine learning, artificial intelligence).[1] By the 1990s, data science arose as a term and later a movement to marshal the combined skills of statisticians and computer scientists to address the study of large datasets. Kelleher and Tierney suggest that by 2001, data science had moved beyond the province of statisticians and computer scientists given increasing heterogeneity in data production (mobile phones, ubiquitous internet access), which required a broader range of technical skills (e.g., scraping, merging, and cleaning data), domain expertise, and the ability to communicate findings to multiple communities.[2]

Machine learning refers to a field of study and practice focused on the development of algorithms—processes or sets of rules—to be followed by a computer to find patterns in data. Patterns discerned within data are represented as models taking a variety of forms (decision trees, neural networks, regression models). Once a model has been produced it can be used to support further analysis through actions like labeling and classifying data (e.g., email spam filters).[3] At a basic level, machine learning is either supervised or unsupervised. Supervised learning refers to a process in which an algorithm is designed to learn how to develop a function that finds specific things in a dataset (e.g., a car, or not a car in an image). Preference toward developing a particular kind of function is referred to as a learning bias. The algorithm is supervised because the dataset is labeled with input and output (target) values that guide the algorithm as it works to develop models that produce the desired function (e.g., finding trees in images). Producing a labeled dataset that supports supervised learning requires a great deal of time and effort. Notably, library metadata can often be used as a labeled dataset. Unsupervised learning works without a labeled dataset to guide it. In this approach, the algorithm looks for regularities in the data (e.g., clustering and re-clustering data that appear to be similar to each other within a dataset). We commonly see the product of unsupervised learning in reading, listening, and viewing recommender systems.

Artificial intelligence (AI) is a field of study and practice that combines methods and areas of focus that include, but are not limited to, natural language processing, machine learning, computer vision, robotics, philosophy, mathematics, neuroscience, psychology, computer engineering, and linguistics in order to create "intelligent machines."[4] In everyday life, AI makes things like voice recognition (Apple's Siri, Microsoft's Cortana, Amazon's Alexa) and image recognition (suggested object tagging for images, facial recognition) possible. AI also supports the development of things like autonomous vehicles (Tesla's attempts to develop self-driving cars), fraud detection systems (Capital One's Eno), robots (Roomba), and efficiencies in supply chain management.[5]

The birth of AI is generally traced to the *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (1955). Since that time, continued effort has been made to create "intelligent machines"; debate as to what features constitute intelligence remain unresolved. Stuart Russell, Peter Norvig, and Ernest Davis argue that, starting in 1987, AI

reached an important state of maturity wherein the field started coalescing around shared theories and committed to rigorous scientific investigation that put proposed advancements in conversation with a wider field of study. For example, much of the work on neural nets in the 1980s was exploratory and it took some time before research methodology was refined to a state wherein it became possible to compare neural nets with corresponding techniques from statistics, pattern recognition, and machine learning. Contemporary approaches to AI are simultaneously catalyzed by the availability of large datasets (e.g., massive digital libraries; image, audio, and video content at web scale) and access to cloud computing.

NOTES

1. Kelleher, John D., and Brendan Tierney. 2018. *Data Science*. The MIT Press Essential Knowledge Series. Cambridge, Massachusetts: The MIT Press.

2. Ibid.

   For an alternative narrative that traces the development of data science further into the past see Donoho, David. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26 (4): 745–66. https://doi.org/10.1080/10618600.2017.1384734.

3. Kelleher et al. *Data Science*. (see note 1)

4. Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Upper Saddle River: Prentice Hall.

5. Toh, Allison. 2018. "AI in Your Wallet: Capital One Banks on Machine Learning." *The Official NVIDIA Blog*. October 12, 2018. https://blogs.nvidia.com/blog/2018/10/12/ai-in-your-wallet-capital-one-banks-on-machine-learning/.

   Snow, Jacob. 2018. "Supply Chain Case Studies." 2019. August 1, 2019. https://www.ibm.com/supply-chain/case-studies.

For more information about our work please visit our website at:
oc.lc/research

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 978-1-55653-152-1
DOI: 10.25333/xk7z-9g97
RM-PR-216345-WWAE 1911