Review Journal of Education and Social Science Research
Vol. 7, No. 8 - August, 2025
ISSN (Print): 2447-6546
ISSN (Online): 2437-3594
https://iigdpublishers.com/

Black Futures: Collecting Sociocultural Data through Machine Learning

Kymberly Keeton[1]
[1]Department of Information Science, University of North Texas, Texas, USA.

Abstract: *Although African American community archives have appeared, there is a lack of incorporation of information-seeking, behaviour, language transmission, categorization, and community archival datasets in data collection and machine learning (ML) environments. To address this, as the autoethnographer, I propose to develop a future body of research targeting the African American community in Texas, obtaining valuable insights about their engagement with ML. Eun Seo Jo's literature review emphasizes the roles of community archives in ML environments and the strategies necessary for this space to be considered a valuable resource in research and information. As the auto-ethnographer, I use this research to explore effective strategies for machine learning environments to collaborate with African American community archives and incorporate user input into ML data collection practices. The aim of the study is to examine an original body of literature to aid me with my plan of action in creating a research study about machine learning in African American community archives.*
Keywords: *African American Community Archives, Sociocultural Issues, Sociocultural Data, Machine Learning, Archives, Datasets, ML Fairness.*

1. Introduction

The importance of African American community archives in the data collection process within machine learning environments has been overlooked by the archival profession. Eun Seo Jo's (2020) article, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning", is a study that highlights the need for Machine Learning (ML) communities to develop authentic approaches for collecting sociocultural data. A significant amount of general literature on data collection in archival studies and valid policies that protect the archive of a subject or artifact on an institution's virtual platform or repository is available to read about. However, the ML community lacks authenticity in their data-collecting practices relating to specific ethnic groups, according to the researcher's study and my observations as the auto-ethnographer of this study. Therefore, the author suggests that ML collectives develop frameworks and strategies for data collection and annotation. The goal of the research study is to examine an original body of literature to help with my plan of action in creating a research study about machine learning in African American community archives.

Definitions

African American or Black: Refers to people who were born in the United States and have African ancestry. It caught on in the U.S. in the 1980s as a more "particular and historical" term than the generic "Black." The terms Black and African American are used interchangeably from a global perspective. Black refers to dark-skinned people of African descent, no matter their nationality. Refers to people who were born in the United States and have African ancestry" (Race Terminology, 2020).

Archives

An archive is a repository that houses primary source materials, such as letters, reports, and photographs, which people can use to gather firsthand facts, data, and evidence (SAA Dictionary: Archives, n.d.).

Artificial Intelligence

Artificial intelligence (AI) refers to computer systems that can perform tasks that traditionally require human intelligence, such as speech recognition, decision-making, and pattern identification (Coursera: What Is Artificial Intelligence? 2023).

Auto-ethnography

"Auto-ethnography is a research method that uses personal experience ("auto") to describe and interpret ("graphy") cultural texts, experiences, beliefs, and practices ("ethno")" (Adams et al., 2017).

Community Archives

"Documentation of a group of people that share common interests, and social, cultural, and historical heritage, usually created by members of the group being documented and maintained outside of traditional archives" ("SAA Dictionary: Community Archives," n.d.).

Machine Learning

A data-oriented technique that enables computers to learn from experience. Human experience comes from our interaction with the environment. For computers, experience is indirect. It is based on data collected from the world, data about the world (Desjardins-Proulx, 2013).

Sociocultural

Sociocultural is related to the diverse groups of people in society and their habits, traditions, and beliefs ("Cambridge Dictionary: Sociocultural" 2023).

2. Review of Literature

In machine learning (ML), the data configurations prompted by its systems play a major factor in the outcomes of research and information. At random, the ML system and its training models can categorically arrange people into unprotected sets and leave them open to societal biases. There is a plethora of tools used today for surveillance and targeting individuals who are not among the elite with access to robust data models. For example, an independent film released on Netflix and PBS' Independent Lens entitled Coded Bias (2021) by Shalin Kantayya is about M.I.T. Media Lab Computer Scientist Joy Buolamwini's discovery that "some algorithms could not detect dark-skinned faces or classify women with accuracy"(Kantayya, 2021). The documentary film intends to show the racial disparities presented within artificial intelligence (AI) and expose prejudices and threats to civil liberties in facial recognition algorithms.

In this case, the researcher explains that there must be diverse data trained to aggregate the data retrieved in instances mentioned through disaggregated examination, i.e., testing. As an example, through said testing, the demographics of communities about identity (i.e., race) strictly use the knowledge of the tester as data. The scholar suggests that in certain instances, though individuals prefer to be affiliated with their choices of representation, there may be times in which they may be affiliated with perceptions based on society or environment (e.g., documenting racial issues during the Georgia elections: African American Democrats flip the state of Georgia in January 2021). For this reason, accurately assessing and developing a cohesive resource for use in ML for systematic processes will include data collections studied by archivists and information professionals for their validity and standards to begin working towards best practices in machine learning environments as they relate to communities.

These findings suggest that there needs to be an interdisciplinary approach to taking on these issues. Eun Seo Jo (2020) insists that the ML community take part in these lessons from other fields such as the archival profession and their long history of striving to always be its data collection methods with integrity and authenticity. The researcher hopes that the ML community will take heed and develop an interdisciplinary subfield that is the implementation of record-keeping processes, annotation, ethics monitoring, data collection, and sharing information. The scholar insists, "We frame our findings about archival strategies into five main topics of concern in the fair ML community: consent, inclusivity, power, transparency, and ethics and privacy" (Jo, 2020).

Differences Between Archival and ML Data Sets

The researcher employs a brief assessment of the two entities and weaves together a tabulation of information that suggests ML practitioners take the approach of looking at the diversity

of data accumulated in collecting data and situating them with the proper vocabulary, which in turn will be its communication strategies. One issue mentioned by the researcher is intervention and supervision in machine learning environments. In theory, reviewing the ML landscape of collecting information, there are no set guidelines, which in turn produce unauthenticated data in mass. Whereas in archival curatorial practices, everything archived is based on the theory of archives. This verbiage allows archivists to filter out what fits within the scope of the archive about a community. Scholarships are available to read about race, machine learning, and artificial intelligence by archivists, and this is upward mobility within the scope of the archival profession, thinking freely with a cause regardless of internal or external politics and race. Thus, to say the least, the archival field has taken on an interventionist approach to archiving history and documents in general.

Contrary to expectations, both the archives and ML fields have different approaches to documenting information. They have distinct areas of expertise and compartmentalize their work accordingly. ML datasets are quickly tabulated to produce information instantly, while archives prioritize categorizing information as metadata and ensuring factual data is embedded in databases or repositories. The archival profession takes terms like the rarity of sources, privacy inclusivity, and authenticity seriously as a practice. It is hypothesized that ML communities should also prioritize these aspects to enhance the validity of their profession and further their aims.

Interventionist Collections Needed in ML

If ML communities do not examine the methods of other fields and develop strategies to supervise and confirm their ethics, biases will persist as information is filtered, potentially distorting reality. ML often reflects historical biases, such as racial disparities, in data, particularly in language data from Asian and Latino communities. Representational bias arises from limitations in digital tools for preservation and digitization. When using natural datasets like web crawling, it is necessary to adopt an interventionist approach to inform participants about the use of their data, even if they can access information online without explicit permission.

Humans may filter information, but that does not guarantee its safety or accuracy. Eun Seo Jo examines how digital news cycles, such as those of the BBC, CNN, and Reuters, intentionally categorize information to align with a political party's ideology or news biases. These media giants do not use data about their audiences to shape ML models but rather aim to cater to their audiences with what they consider news. Eun Seo Jo (2020) suggests that machine learning researchers can assert their authority by using data that is not safeguarded by these news organizations. The scholar emphasizes the need for critical investigations into the purpose, objectives, and practices of these data sets, advocating for an interventionist data collection approach.

Inclusivity in ML: Mission Statements & Collection Policies

Eun Seo Jo (2020) emphasizes the need for ML and archives to create mission statements that support the representation of cultural artifacts and diverse demographics. It is crucial to regularly assess and address biases in these fields to promote diversity, equity, and inclusion.

Approval: Community & Participatory Archives

Anglo institutions have been gathering records since the 1970s about Black, Asian, and Latino communities. Today, these records are known as community archives and consist of papers, notes, memorabilia, photos, and anything that stands for the identity of the said community. The purpose of these archives is to ensure representation for marginalized, grassroots and none lite communities. In the 21st century, multiple digital and self-collecting archival institutions represent various gender, cultural, educational, artistic, and religious communities.

Community archives serve as a means for public collaboration in the democratization of collecting practices in machine learning, granting autonomy to these communities to own their voices and data and represent themselves. Additionally, this empowers minority groups to show their classification systems with consent. The author of this article has not supplied any details about African American community archives and the necessary strategies for representing the collecting practices of the ML field in terms of sociocultural data, knowledge acquisition, archival categorization, and language transmission. All to say, there is a gap in the literature and research about this subject.

Power: Data Consortia

Ethics in data collection requires a significant amount of time to define in any organization. In the field of ML, there is no accountability for the tabulated data. To promote balanced data ownership, archives and libraries collaborate in consortia spaces and set up best practices. Consortia, as defined in this research, are shared resources, collection networks, and distribution services known as library networks or cooperatives (p. 311). Jo lists the consortia members as AMIGOS Library Services, MELSA, and OCLC (p. 311). The main goal of consortia is to achieve economies of scale by collaborating on massive projects that can be shared and archived in various spaces less than one server. This collaboration helps offset costs for repository space and data storage as it is shared among the consortiums.

Archives and libraries face criticism for their use of consortia. These consortia have set up bureaucratic committees that cause delays in releasing collections and are seen as elitist. They are funded through memberships and offer lucrative benefits and power. In the field of machine learning, scholars recognize the inseparable connection between economic profit and data. The ML community in the UK is currently working towards greater collaboration in data sharing, but no information is given about the situation in the United States.

Transparency Appraisal Flows with Data Collection

A key factor in the ML community now being at the forefront of their discussions is the lack of communication that is portrayed in this environment about data collection and ML model architecture. For example, the researcher states that in Datasheets for Datasets, archivists can break down inquiries that require scholars to describe how they retrieved a particular dataset. Acknowledging and implementing a directive measure that is given in the data-collection process allows for transparency. The archival profession has standards for data descriptions, and it is the job of the archivist to abide by these standards from the author's scholarship:

Data Content Standards
Specifies the syntax of data, order, and content.
Data Structure Standards
Specifies the organization of data (EAD, EAC-CPF)
Data Value Standards
Specifies the terms used to describe data (LCSH, AAT, NACO)

Data Appraisal Flow

The study presents an example of appraisal processes in archives that advance through levels of supervision by archivists, curators, and records managers, offering various explanations for the standards involved by the author:

Mission Statement
Agenda documentation that provides topics and concepts of concern

Collection Development Policy
Adapted from the mission statement about what is collected and what is not and where and how to search for sources.

Appraisal
Evaluation criteria based on a selection of sources worthy of being included in the collection.
Questioning the validity of the source and if it is in line with the mission statement.
Evaluate the authenticity of collected data.

Processing/Indexing (Micro-Appraisal)
Processing and indexing collections or data.
Sources can be discarded due to privacy.
These models noted are strategies for the ML community to adapt to their professional vernacular.

Codes of Conduct: Ethics and Privacy

The Society of American Archivists, through their website, states that their Core Values for Archivists and Code of Ethics for Archivists are meant to be used together as a guide for individuals

working in archival settings. These values and principles shape the expectations for professional behavior and involvement within the field. The organization aims for authenticity in the use and documentation of archival materials, promoting access for all communities, ensuring accountability, and preserving collections for future generations (SAA Core Values Statement and Code of Ethics | Society of American Archivists, n.d.). By enhancing its ethical framework and incentivizing adherence to authoritative standards, the ML industry can surmount its existing obstacles. This can be accomplished through the advocacy of permanent positions in data collection, facilitating the enforcement of ethical guidelines, and ensuring accountability among data collectors (p. 313).

Actionable: Two Levels

Using the following action levels by Eun Seo Jo can improve ML data collection and annotation:

Macro Level

Community, private institutions, policymakers, and government agencies
Organize and create data consortia.
Develop professional spaces working by membership to enforce ethical guidelines.
Advocate for community archives.

Micro Level

Individual researchers, practitioners, and administrators
Define and create mission statements.
Employ full-time data collectors, curators, and administrators whose performance is aligned with the macro and micro levels.
Document and adopt standards.
Create a collection development policy and update the domain and nuance of data sources regularly.
Make committee decisions about cautionary data.

Critical Observation: Limitations in Spaces

An important issue emerging from these findings is that ML datasets are large in scope and require more studies on how they are transferable in the context of machine learning in data science. For example, being able to employ a full-time worker, store documentation, and implement collection development strategies is costly regarding large-scale data collection. As claimed by the researcher (2020), who was earlier affiliated, "Maintaining data consortia at the community level is one way to reduce these costs through economies of scale, resource sharing, and minimizing duplicity" (p. 315).

The main limitation of Jo's study is interventionist models and how they too embody selectivity in data collection practices based on the biases of archivists, who have the autonomy to make decisions about collections and treat them unethically. Another example the scholar provides is the influx of political and social ideologies that systematically harbor exclusion in certain communities. The main weakness in this study is that the author does not necessarily discuss the lack of diversity in depth regarding race and gender in ML as a professional or advocacy for jobs to be created for individuals from ethnic communities.

3. Conclusion

Throughout my research and study of the literature reviewed, I was compelled to shift and take a leap toward a new focus on African American community archives and their relationship with machine learning. This area of exploration holds significant importance due to the lack of scholarly attention it has received. Machine learning has the potential to impact decision making processes in libraries, archives, and museums, which can often lead to the exclusion of certain communities. I commend Eun Seo Jo for his transparent and inclusive approach to addressing the importance of archival work as a practice within the scope of machine learning and artificial intelligence in his article.

This study contributes to the growing body of knowledge on the prominence of machine learning in our daily lives and the sociocultural dynamics that have and will arise in the archival profession. However, for this knowledge to hold value, the machine-learning community must prioritize the macro- and micro-levels of action outlined by the scholar. This will ensure that profit-driven agendas do not monopolize the field. The study highlights the significance of archival practices

in data collection, such as collection development policies, ethics standards, and mission statements. As the auto-ethnographer, I show that these practices are well within the scope of consideration for the machine learning community. In conclusion, my research has shed light on the societal biases that have led to gaps in African American archives and data science. It is crucial to address these biases to ensure that the machine-learning community is inclusive and representative of all communities.

Black Futures Reflection: Future Objectives of Study

Abstract

As the auto-ethnographer conducting this study, my primary motivation for selecting this topic stems from my curiosity and passion for African American community archives. I am particularly fascinated by the potential of machine learning in community archiving, specifically within African American communities. The preservation of their history through categorical information is of great interest to me. Through my research, I ask the question: Are African Americans in Texas actively using machine learning as an indispensable tool with community archives? To further delve into this topic, I intend to construct a plan of action using a mixed methods approach to explore the use of machine learning in the Black community. This research, as a future practice, will serve as a foundation for a forthcoming research project and scholarship.

Definitions

Auto-ethnography
African American/Black
Machine Learning
Artificial Intelligence
Community Archives
Sociocultural Data

Research Question

Are African Americans in Texas actively using machine learning as an indispensable tool with community archives?

Literature Review Bibliography

Non-exhaustive Bibliographic Listing of Articles

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 289–298. https://doi.org/10.1145/3287560.3287575

Stevenson, J. (2022). Machine Learning with Archive Collections – Archives Hub Blog. Machine Learning with Archive Collections. https://blog.archiveshub.jisc.ac.uk/2022/02/28/machine-learning-with-archivecollections/

Prescott, A. (2023). DHQ: Digital Humanities Quarterly: Bias in Big Data, Machine Learning and AI: What Lessons for Digital Humanities? Digital Humanities Quarterly, 17(2). https://digitalhumanities.org/dhq/vol/17/2/000689/000689.html

Terras, M. (2022). Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. In Archives, Access, and AI: Working with Born-Digital and Digitized Archival Collections (pp. 179–204). https://doi.org/10.1515/9783839455845-008

Taurino, G., & Smith, D. A. (2022). Machine Learning as an Archival Science: Narratives behind Artificial Intelligence, Cultural Data, and Archival Remediation. Contributed Panel 2b: Theorizing AI/Culture Entanglement. A NeurIPS 2022 Workshop, Virtual.

Wiggers, K. (2020). Could machine learning help bring marginalized voices into historical archives? VentureBeat. https://venturebeat.com/business/could-machinelearning-help-bring-marginalized-voices-into-historical-archives/

I Exert: Theory

For my research, I have formulated a plan to use a mixed-methods approach that incorporates a diverse range of theories from both information science and the social sciences. Specifically, I will be drawing upon the following theories: Archival (Machine Learning), which applies machine learning techniques to archive management and organization; Black Archival Practice, which explores

the role of archives in preserving and promoting black cultural heritage; Social Closure, which examines how groups restrict access to resources and information; and Community Archives Theory, which is about the importance of community engagement in archive creation and management. Lastly, I will divulge Sociocultural Perspective Theory, used to describe awareness of circumstances surrounding individuals and how their behaviors are affected specifically by their surroundings and social and cultural factors. By using these theories, I hope to gain a comprehensive and nuanced understanding of the complex issues surrounding archival practices and their impact on society.

4. Methodology

The guiding approach and method introduced for this study are embedded in auto-ethnography, which involves placing oneself at the center of an archival analysis. In this case, machine learning is a tool for community archiving. As the auto-ethnographer, I use the method of participatory action research (survey and virtual forum) to examine the use of ML through language, culture, knowledge, and visual frames. Action research is defined by Anne Burns (2015) as "a set of research approaches that, at the same time, systematically investigate a given social situation and promote democratic change and collaboration participation" (Burns, 2015, p. 187). Utilizing this method offers a myriad of benefits, as it grants researchers the freedom to conduct interviews either in person, virtually, over the phone, or through mail. As a result, participants can genuinely convey their thoughts on the topic of interest while allowing researchers to analyze and classify data in multiple ways, thereby helping the visualization process and generating valuable insights.

Proposed Instruments

For this research, I will apply for an IRB-approved research study survey that will be available as a virtual survey and online forum with the following criteria:
Participants: Ages 18 and beyond.
Target Group: African Americans in Texas and or who identify as Black
Recruitment: Poster, Email, Facebook, LinkedIn, and mobile devices.

Sample Survey Questions

Define data.
Define community archives.
Define sociocultural.
Define race.
Are you familiar with the term Machine Learning?
Are you familiar with the term Artificial Intelligence?
Have you ever enrolled in a course for ML or AL?
Have you participated in a coding project using Python?
Have you ever used ML or AI as a genealogy model?
Have you ever used ML or AI for a community archive model?
Have you ever used ML or AI to change an image or artwork?
Have you ever used text mining before?
Do you plan to use ML or AL in your professional work life?
What has your experience been like using Machine Learning as a tool?

Data Collection

To collect the necessary data, participants will have the choice between completing a survey questionnaire or taking part in a virtual forum interview. This choice must be made before giving their responses to the study. The survey questionnaire will consist of a set of questions on the relevant topic, while the virtual forum interview will be conducted through a digital platform where participants can share their thoughts and opinions in a group discussion setting. Both methods are designed to gather valuable insights and feedback from the participants.

Data Analysis

As part of the data analysis assessment, the collected data will be meticulously examined and manually inputted into IBM SPSS software. Using the software, we will generate comprehensive reports that have crucial demographic information and insights derived from the answers provided in

the questionnaire. To analyze the forum feedback, we will use text mining by taking quotes from participants and entering them into a corpus that will supply significant text analysis based on language redundancy, commonality, and deviation.

5. Limitations

This study's constraints lie in the fact that the participants are not sufficiently informed about machine learning, a subdivision of artificial intelligence. Additionally, obtaining enough participants who are willing to engage in the forum and share their experiences may prove challenging due to their limited understanding of machine learning and archives.

6. Summary

As an experienced researcher and practitioner in this field, I have thoroughly examined the evidence gathered so far and concluded that it is compelling enough to justify continuing with this project and study. Based on my analysis, I am confident that the data obtained will help me gain a deeper understanding of the subject matter and potentially lead to valuable insights and solutions. Therefore, I recommend that this study (as the auto-ethnographer) continue with the next phase of this research with a clear focus on achieving the desired outcomes.

References

Adams, T., Ellis, C., & Jones, S. (2017). Autoethnography. https://doi.org/10.1002/9781118901731.iecrm0011

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 289–298. https://doi.org/10.1145/3287560.3287575

Burns, A. (2015). Action research. In The Cambridge Guide to Research in Language Teaching and Learning (pp. 99–104). Cambridge: Cambridge University Press.

Coursera. (2023). What Is Artificial Intelligence? Definition, Uses, and Types. Coursera. https://www.coursera.org/articles/what-is-artificial-intelligence

Devopedia. (2020). "Machine Learning." Version 18.

Jo, E. S., & Gebru, T. (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 306–316. https://doi.org/10.1145/3351095.3372829

Kantayya, S. (Director). (2021). Coded Bias | Films | PBS. Independent Lens. https://www.pbs.org/independentlens/films/coded-bias/

Prescott, A. (2023). DHQ: Digital Humanities Quarterly: Bias in Big Data, Machine Learning and AI: What Lessons for Digital Humanities? Digital Humanities Quarterly, 17(2). https://digitalhumanities.org/dhq/vol/17/2/000689/000689.html

SAA Core Values Statement and Code of Ethics | Society of American Archivists. (n.d.). https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics

SAA Dictionary: Community archives. (n.d.). https://dictionary.archivists.org/entry/community-archives.html

Sociocultural. (2023). In Cambridge Dictionary. Cambridge University Press & Assessment. https://dictionary.cambridge.org/dictionary/english/sociocultural

Stevenson, J. (2022). Machine Learning with Archive Collections – Archives Hub Blog. Machine Learning with Archive Collections. https://blog.archiveshub.jisc.ac.uk/2022/02/28/machinelearning-with-archive-collections/

Taurino, G.,& Smith, D. A. (2022). Machine Learning as an Archival Science: Narratives behind Artificial Intelligence, Cultural Data, and Archival Remediation. Contributed Panel 2b: Theorizing AI/Culture Entanglement. A NeurIPS 2022 Workshop, Virtual.

Terras, M. (2022). Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. In Archives, Access, and AI: Working with Born-Digital and Digitized Archival Collections (pp. 179–204). Transcript Verlag. https://doi.org/10.1515/9783839455845-008

The University of Central Arkansas, Training Files, (2020). Race Terminology. SAA Dictionary: Archives, (n.d.). https://dictionary.archivists.org/entry/archives.html

Wiggers, K. (2020). Could machine learning help bring marginalized voices into historical archives? VentureBeat. https://venturebeat.com/business/could-machine-learning-help-bringmarginalized-voices-into-historical-archives/