Chapter 12

Machine Learning + Data Creation in a Community Partnership for Archival Research

Jason Cohen
Berea College

Mario Nakazawa
Berea College

Introduction: Cultural Heritage and Archival Preservation in Eastern Kentucky

In this chapter, two researchers, Jason Cohen and Mario Nakazawa, describe the contexts for an archivally focused project that emerged from a partnership between the Pine Mountain Settlement School (PMSS)[1] in Harlan County, Kentucky, and scholars and students at Berea College. In this process, we have entered into a critical dialogue with our sources and knowledge production that Roopika Risam calls for in "self-reflexive" investigations in the digital humanities (2015, para. 16). Risam's intervention, nevertheless, does not explicitly distinguish questions of class and the concomitant geographic constraints that often accompany the economic and social disadvantages of poverty (Ahmed et al. 2018). Our work demonstrates how class and geography are tied, even in digital archives, to the need for reflexive and diverse approaches to humanist materials. For instance, a recent invited contribution to *Proceedings of the IEEE* articulates a need

[1] See `http://pinemountainsettlementschool.com`.

for diversity in computing and technology without mentioning class or region as factors shaping these related issues of diversity (Stephan et al. 2012, 1752–5). Given these constraints, perhaps it is also pertinent to acknowledge that the machine learning application we describe in this chapter is itself not particularly novel in scope or method—we describe our data acquisition and preparation, and two parallel implementations of commercially available tools for facial recognition. What stands out as unique are the ethical and practical concerns tied to bringing unique archival materials out of their local contexts into a larger conversation about computer vision as a tool that helps liberate, and at the same time possibly endanger, a subaltern cultural heritage.

In that light, we enter our archival investigation into what Bruno Latour has productively named "actor-network theory" (2007, 11–13) because, as we suggest below, our actions were highly conditioned not only by the physical and social spaces our research occupies and where its events occurs, but also because the nature of the historical artifacts themselves act powerfully to shape our work in these contexts. Moreover, the partnership model of curation and archiving that we pursued in this project complicates the very concept of agency because the actions forming the project emerged from a continuing dialogue rather than any one decision or hierarchy. As we suggest later, a distributed model for decisions (Sabharwal 2015, 52–5) also revealed the limitations of using a participatory and identity-based model for archival development and management. Indeed, those historical artifacts will exert influence on this network of relations long after any one of us involved in the current project has ceased to pursue them. When we came to this project, we asked a version of a classic question that has arisen in a variety of forms beginning with very early efforts by Bell Laboratories, among others, to translate data structures to suit the often flexible needs of humanist data: "what aspects of life are formalizable?" (Weizenbaum 1976, 12). We discovered that while an ontology may represent a formalized relationship of an archive to a database or finding aid, it also asks questions about the ethical implications of what information and embedded relationships can be adequately formalized by an abstract schema.

The Promises and Realities of Technology After Coal in Eastern Kentucky

Despite the longstanding threats of having to adapt to a post-coal economy, Harlan County, Kentucky continues to rely on coal and the mountains from which that coal is extracted as two of the cornerstones that shape the identity of the territory as well as the people who call it home. The mountains of Eastern Kentucky, like much of Appalachia, are by turns beautiful and devastated, and both authors of this essay have found conversations with Eastern Kentucky's citizens about the role the mountains play and the traditions that emerge from them both insightful and, at times, heartbreaking. This dramatic landscape, with its drastic challenges, may not sound like a place likely to find uses for machine learning. You would not be alone in your assumption.

Standing far from urban centers of technology and mobility, Eastern Kentucky combines deeply structural problems of generational poverty with a hard won understanding that, since the moment of the region's colonization, outsiders have taken resources and made uninformed decisions about what the region needs, or where it should turn in order to gain a better purchase on the narrative of American progress, self-improvement, and the unavoidable allures of development-driven capitalism. Suspicion of outsiders is endemic here. And unfortunately, economic and social conditions, such as the high workplace injury rates associated with mining and extraction-related industries, the effects of the pharmaceutical industry's abuse of prescription

opioids to treat a wide array of medical pain symptoms without treating the underlying causal conditions, and the systematic dismantling of federal- and state-level social support programs, have become increasingly acute concerns today. But this trajectory is not new: when President Lyndon B. Johnson announced the beginning of the War on Poverty in 1964, he landed an hour away in Martin County, and subsequently, drove through Harlan on a regional tour to inaugurate the initiative. Successive generations have sought to leave a mark, and all the while, the residents have been collecting their own local histories of their place. Our project, centered on recovering a latent social network of historical families represented by the images held in one local archive, mobilizes this tension between insiders' persistence and outsiders' interventions to think about how, as Bruno Latour puts it, we can "reassemble the social" while still respecting the local (2007, 191–2). PMSS occupies a unique position in this social and physical landscape: both local in its emplacement and attention, and a site of philanthropic work that attracted outside money as well as human and cultural capital, PMSS is at once of Harlan County and beyond it. As we suggest in the later sections of this essay, PMSS's position, both within local and straddling regional boundaries, complicates the network we identified. More than that, however, its split position complicates the relationships of power and filiation embedded in its historical social network.

While an economy centered on coal continues to define the Eastern Kentucky regional identity, a second history can be told about this place and its people, one centered on resilience, independence, simplicity, and beauty, both of the land and its people. This second history has made outsiders' recent appeals for the region to court technology as a potential solution for what comes "after coal" particularly attractive to a region that prides itself on its capacity to sustain, outlast, and overcome obstacles. While that techno-utopian vision offers another version of the self-aggrandizing Silicon Valley bootstraps success story J.D. Vance narrates in *Hillbilly Elegy* (2016), like Vance's story itself, those narratives most often get told by outsiders to outsiders using regional stereotypes as the grounds for a sales pitch. In reality, however, those efforts have largely proven difficult to sustain, and at times, become the sources of potentially explosive accusations of fraud and malfeasance. Recently, for instance, organizations including Mined Minds[2] have been accused by residents aiming to prepare for a post-coal economy of misleading students, at least, and of fraud at worst. As with the timber, coal, and gas extraction industries that preceded these software development firms' aspirations, the promises of technology have not been kind to Eastern Kentucky, and in particular, as with those extraction industries that preceded them, the technological-industrial complex making its pitch in Kentucky's mountains has not returned resources to the region's residents whom the work was intended at least nominally to support (Hochschild 2018; Campbell 2019; Bailey 2017).

In this context of technology, culture, and the often controversial position machine learning occupies in generating obscure metrics for its classifiers that may embed bias, our project aims to activate its archival holdings and bring critical awareness to the question of how to actively engage with a paper archive of a local place as we venture further into our pervasively digital moment. The School operates today as a regional cultural heritage institution; it opened in 1913 as a residential school and operated as an educational institution until 1974, at which point it transformed itself into an environmental and cultural outreach institution focused on developing its local community and maintaining the richness of the region's cultural resources and heritage. Every year since 1974, PMSS has brought hundreds of students and citizens onto its campus to learn about nature and the landscape, traditional crafts and artistic practices, and musical and dance forms, among many other programs. Similarly, it has created a space for locals to come

[2] See http://www.minedminds.org/.

together for social events, community celebrations, and festival days, and at the same time, has become a destination for national-level events that create community from shared interests including foodways, wildflowers, traditional dance forms, and other wide-ranging attractions.

Project Background: Preserving Cultural Heritage in Harlan Country

The archives of the Pine Mountain Settlement School emerge from its shifting history. The majority of its papers relate to its time as a traditional institution of education, including student records (which continue to be restricted for several reasons, including FERPA constraints, and personal and community interests in privacy), minutes of its board meetings (again, partially restricted), and financial and narrative accounts of its many activities across a year. The school's records are unique because they provide a snapshot, year by year and month by month, of the region's interests and challenges during key years of the 20th Century, spanning the First World War to Vietnam. In addition, they detail the relations the School maintained with a philanthropic base of donors who helped to support it and shape it, and beyond its local relations, place it into contact with a larger set of cultural interactions than a boarding school that relied on tuition or other profit-driven means to sustain its operations would. While the archival holdings continued to be informally developed by its directors and staff, who kept the official papers organized roughly by year, the archive itself sat largely neglected after 1974. Beginning around the turn of the millennium, a volunteer archivist named Helen Wykle began digitizing items one by one, and soon, hosted a curated selection of those digital surrogates along with interpretive and descriptive narration on a WordPress installation, The Pine Mountain Settlement School Collections.[3] The PMSS Collections WordPress site has been continuously running and frequently updated by Wykle and the volunteer community members she has organized since 1999.[4] Together with her collaborators and volunteers, Wykle has grown the WordPress site to over 2200 pages, including over 30,000 embedded images that include photographs and newspapers; scanned memos, meeting minutes and other textual material (in JPG and PDF formats); HTML transcriptions and bibliographies hard-coded into the pages; scanned images of 3-D collections objects like textile looms or wood carving tools; partially scanned runs of serial publications; and other composite visual material. None of those objects was hosted within a regular and complete metadata hierarchy or ontology: no regular scheme of fields or file-naming convention was followed, no controlled vocabulary was maintained, no object-types were defined, no specific fields were required prior to posting, and perhaps unsurprisingly as a result, the search and retrieval functions of the site had deteriorated noticeably.

In 2016, Jason Cohen approached PMSS with the idea of using its archives as the basis for curricular development at Berea College.[5] Working in collaboration beginning in 2017, Mario Nakazawa and Cohen developed two courses in digital and computational humanities, led a team-directed study in augmented reality in coordination with Pine Mountain, contributed ma-

[3] See `https://pinemountainsettlement.net/`.

[4] Jason Cohen and Mario Nakazawa wish to extend a note of appreciation to Helen Hays Wykle, Geoff Marietta, the former director of PMSS, and Preston Jones, its current director, for welcoming us and enabling us to access the physical archives at PMSS from 2016–20.

[5] Jason Cohen would like to recognize the support this project received from the National Endowment for the Humanities' "Humanities Connections" grant. See grant number AK-255299-17, description online at `https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=AK-255299-17`.

terials and methods for a new course in Appalachian Studies, and promoted the use of PMSS archival materials in several other extant courses in history and art history, among others. These new college courses each make use of PMSS historical documents as a shared core of visual and textual material in a digital and computational humanities concentration that clusters around critical archival and textual studies.[6]

The success of that initial collaboration and course development seeded the potential in 2019–2021 for a Whiting Public Engagement[7] fellowship focused on developing middle and high school curricula for use in Kentucky public schools with PMSS archival materials. That Whiting funded project has generated over 80 lessons keyed to Kentucky state standards; these lessons are currently in use at nine schools across eight school districts, and each school is using PMSS materials to highlight its own regional and local interests. The work we have done with these archives has thus far reached the classrooms of at least eleven different middle and high school teachers, and as a result, touched over 450 students in eastern and central Kentucky public schools.

We mention these numbers in order to demonstrate that our collaboration has not been shallow nor fleeting. We have come to know these archives quite well, and because they are not adequately cataloged, the only way to get to know them is to spend time reading through the materials one page at a time. An ancillary consequence of this durable collaboration and partnership across the public-academic divide is the shared recognition early in 2019 that the PMSS archival database and its underlying data structure (a flat SQL database generated by the WordPress interface) would provide inadequate stability for records management and quality control in future development. In addition, we discovered that the interpretive materials and metadata associated with the WordPress installation were also insufficient for linked metadata across the objects in this expanding digital archive, for reasons discussed below.

As partners, we decided together to migrate to a ContentDM instance hosted by the Kentucky Virtual Library,[8] a consortium to which Berea College belongs, and which is open to future membership from PMSS. That decision led a team of Berea College undergraduate and faculty researchers to scrape the data from the PMSS archive site and supplement the images and transcriptions it contains with available textual metadata drawn from the site.[9] Alongside the WordPress instance as our reference, we were also granted access to a Dropbox account that hosted higher resolution versions of the images featured on the blog. The scraper pulled over 19,228 unique images (and located over 11,000 duplicate images in the process), 732 document transcriptions for scanned texts on the site, and 380 subject and person bibliographies, including Library of Congress Subject Headings that had been hard-coded into the site's HTML. We also extracted the unique object identifiers and labels associated with each image, which in WordPress are not associated with the image objects themselves. We used that data to populate the ContentDM instance and returned a sparse but stable skeleton for future archival development. In the process, we also learned significantly about how a future implementation of a controlled vocabulary, an image acquisition and processing pipeline, and object documentation standards should work in the next stages of our collaborative PMSS archival development.

[6] In the original version of the collaboration, we had planned also to teach basic computer programming to high school students during a summer program that also would have used that same set of materials, but with the paired departures of the original co-PI as well as the former director, that plan has thus far remained unfulfilled.

[7] See `https://www.whiting.org/content/jason-cohen`.

[8] See `https://kdl.kyvl.org/`.

[9] Jason Cohen wishes to thank Mario Nakazawa, Bethanie Williams, and Tradd Schmidt for undertaking this project with him. The github repo for the PMSS scraper is hosted here: `https://github.com/Tradd-Schmidt/PMSS_Scraper`.

As we developed and refined this new point of entry to the digital archives using the ContentDM hosting and framework, some of the ethical issues surrounding this local archive came more clearly into focus. A parallel set of questions arose in response in the first instance to J.D. Vance's work, and in the second, to outsiders' claims for technological solutions to the deterioration of local and cultural heritage. Because we were creating virtual archival surrogates for materials housed at Pine Mountain, for instance, questions arose from the PMSS board members related to privacy and use of historical materials. Further, the board was concerned that even historical materials could bear on families present in the community today. We found that while profession-wide responses to archival constraints are shaped predominantly by discussions of copyright and fair use, issues of personal privacy are often left tacit. This gap between legal use and public interests in privacy reveals how tasks executed using techniques in machine learning may impinge upon more ethical constraints of public trust and civic obligation.[10]

Similarly, as the ownership of historical images suddenly extended to include present-day community members, and as these questions of access and serving a local public were inextricably bound up with interactions with members of that shared public whose family names and faces appear in the images we were making available, we began to consider the ways in which our archival work was tied to what Ryan Calo calls the "historical validation" of primary source materials (2017, 424–5). When an AI system recognizes an object, Calo remarks, that object is validated. But how should one handle the lack of a specific vocabulary within a given training set? One answer, of course, would be to train a new set—but that response is becoming increasingly prohibitive for smaller cultural heritage projects like ours: the time and computational power required to execute the training is non-negligible. In addition, training resources (such as data sets, algorithms, and platforms) are increasingly becoming monetized, and we do not have the margins to buy access to new data for training. As a consequence, questions stemming from how one labels material in a controlled vocabulary were also at issue. We encountered a failure in historical validation when, for instance, our AI system labeled a "spinning wheel" as a wheel, but did not detect its historical relationship to weaving and textiles. That validation was further obscured when the system also failed to categorize a second form of "spinning wheel," which refers locally to a home-made merry-go-round.[11] In other words, not only did the system flatten a spinning wheel into a generic wheel, it also missed the regional homology between textile production and play, a cultural crux that reveals how this place envisions an intersection between work and recreation. By breaking the associations between two forms of "spinning wheel," our system erased a small but significant site of cultural inheritance. How, we asked, should one handle such instances of effacement? At one level, one would expect an archival system to be able to identify the primitive machine for spinning wool, flax, or other raw materials into usable thread for textiles, but what about the merry-go-round? And what should one do when a system neglects both of these meanings and reduces the object to the same status as a wheel on a tractor, car, or carriage?

Similarly, when competing naming conventions arise for landmarks, we were conscious to consider which name should be granted priority as the default designation, and we asked how one should designate a local or historical name, whether for a road, waterway, knob, or other feature, in relationship to a more widely accepted nomenclature such as state route designations or

[10] The professional conversation in archive and collections management has not been as rich as the one emerging in AI contexts more broadly. For a recent discussion of the conflict in the roles of public trust and civic service that emerge from the context of the powers artificial intelligence holds for image recognition in policing applications, see Elizabeth Joh, "Artificial Intelligence and Policing: First Questions," *Seattle University Law Review* 41: 1139–44.

[11] See "Spinning Wheel" in Cassidy 1985–2012.

standardized toponym? As we attempted to address the challenge of multiple naming conventions, we encountered some of the same challenges that archivists find in dealing with indigenous peoples and their textual, material, and physical artifacts.[12] Following an example derived from the Passamaquoddy people, we implemented a small set of "traditional knowledge labels"[13] to describe several forms of information, including (a) restrictions on images that should not be shown to strangers (to protect family privacy), (b) places that should remain undisclosed (for instance, wild ginseng, ramp, orchid, or morel mushroom patches), and (c) educational materials focused on "how it was done" as related to local skills and crafts that have more modern implementations, but for which the traditional practices have remained meaningful. This included cases such as Maypole dancing and festivals, which remain endowed with ritual significance. In the final analysis, neither the framework supplied by copyright and fair use nor the one supplied by data validation proved singularly adequate to our purposes, but they did provide guidelines from which our facial recognition project could proceed, as we discuss below.

Machine Learning in a Local Archive

These preliminary discussions of ethics and convention may seem unrelated to the focus this collection adopts toward machine learning and artificial intelligence in the archive. However, as we have begun to suggest, the data migration to ContentDM opened the door to machine learning for this project, and those initial steps framed the pitfalls that we continue to navigate as we continue forward. As we suggested at the outset, the technical machine-learning task that we set for ourselves is not cutting edge research as much as an application of existing technologies to a new aspect of archival investigation. We proposed (and succeeded with) an application of commercial facial recognition software to identify the persons in historic photographs in the PMSS archives. We subsequently proposed and are currently working to identify the photographs sharing common but unnamed faces, and in coordination with photographs of known people, to re-create the social network of this historic institution across slices of its history.

We describe the next steps briefly below, but let us tarry for a moment with the question of how the ethical concerns we navigated up to this point also influenced our approach to facial recognition. The first of those concerns has to do with commercial and public access to archival materials that, as we suggested above, include materials that are designated as restricted use in some way. We demonstrated to the local members at Pine Mountain how our use case and its constraints for digital archives fit with the current standards for the fair use of copyrighted materials based on the "substantive transformation" of reproduced objects (Levendowski 2018, 622–9). Since we are not making available large bodies of materials still protected by copyright, and since our use of select materials shifts the context within which they are presented, we were able to negotiate with PMSS to allow us to design a system for facial recognition using the ContentDM instance as our image source. What that negotiation did not consider, however, is when fair use does not provide a sufficiently high standard of control for the institution involved in the application of algorithms to institutional memory or its technological dependencies.

First, to test the facial recognition processes, we reached back to the most primitive and local version of facial recognition software that we could find, Google's retired platform, the Picasa

[12] One well-documented digital approach to handling indigenous archival materials includes the Mukurtu platform for indigenous cultural heritage: `https://mukurtu.org/`

[13] For the original traditional knowledge labels, see: `https://passamaquoddypeople.com/passamaquoddy-traditional-knowledge-labels`.

Web Albums API, which was retired in May 2016 and fully deprecated as of March 2018 (Sabharwal 2016). We chose Picasa because it is a self-contained software application that operates using a locally hosted script and locally hosted images. Given its deprecated status and its location on a local machine, we were confident that no cloud services would be ingesting the images we fed into the system for our trial. This meant that we could test small data examples without fear of having to upload an entire corpus of material that could subsequently be incorporated into commercial facial recognition engines or pop up unexpectedly in search results. We thus began by upholding a high threshold for privacy and insisting on finding ways for PMSS to maintain control over these images within the grasp of its local directories.

The Picasa system created surprisingly good results within the scope we allowed it. It was highly successful at matching the small group of known faces we supplied as test materials. While it would be difficult to supply a numerical match rate first because of this limited test set, and second because we have not expanded the test to a broad sample using another platform, we were anecdotally surprised at how robust Picasa's matching was in practice. For instance, Picasa matched the images of a single person's face, Celia Cathcart, from pictures of her as a teenager to images of her as a grandmother. It recognized Cathcart in a group of basketball players, and it also identified her face from side-view and off-center angles, as in a photograph of her looking down at her newborn child. The most immediate limitation of Picasa lies in its tagging, which required manual entry of every name and did not allow any automation.

Following the success of that hand-tagging and cross-image identification process, we discussed with our partners whether the next step, using Amazon Web Services' computer vision and facial recognition platform, ReKognition, would be acceptable. They agreed, and we ran the images through the AWS application, testing our results against samples pulled from our Picasa run to verify the results. Perhaps unsurprisingly, AWS ReKognition fared even better with those test cases. Using one photograph image, the AWS application identified all of the Picasa matches as well as three new images that had not previously been tagged with Cathcart's name. The same pattern held for other images in our sample group: Katherine Pettit was positively identified across more likenesses than had been previously tagged, and Alice Cobb was also positively tracked across images. This positive attribution also reveals a limitation of the metadata: while these three women we have named are important historical figures at PMSS, and while they are widely acknowledged in the archive and well-represented in the photographic record, not all of the photographs have been well-tagged or fully documented in the archive. The newly tagged images that we found would enrich the metadata available to the archive not because these images include surprising faces, but rather, because the tagging has been inconsistent, and over time, previously known faces have become less easy to discern.

Like other recent discussions of private materials disclosed within systems trained for matching and similarity, we found that the ethics of private materials for this non-private purpose provoked strong reactions. While some of the reaction was positive with community members happy to have more images of the School's founding director, Katherine Pettit, identified, those same community members were not comfortable with our role as researchers identifying people in the photographs in their community's archive, unsupervised. They wanted instead to verify each positive identification, a point that we agreed with, but which also hindered the process of moving through 19,000 images. They wanted to maintain authority, and while we saw our efforts as contributions to their goals of better describing their archival holdings, it turns out that the larger scope of automation we brought to the project was intimidating. While its legal status and direct ethics seemed settled before the beginning of the project, ultimately, this project contributed to

a sense among some individuals at PMSS that they were losing control of their own archive.[14] That fear of a loss of control led to another reckoning with the project, as we discuss in the next section.

What Machine Learning Cannot Learn: An Ethics of the Archive

It became clear at the same moment we validated our test case, that our research goals and those of our partners had quickly diverged. We had discussed the scope and use of PMSS materials with our partners at PMSS and laid out in a formally drafted "Memorandum of Understanding" (MOU) adapted from the US Department of Justice (2008; 2017) our shared goals in the project. As we described in the MOU, both partners considered it mutually beneficial for the archive and its metadata to be able to identify faces of named as well as unnamed people. We aimed to capture single-person images as well as groups in order to enrich the archive with cross-links to other photographs or archival materials with a shared subject heading, and we hoped to increase the number of names included in object attributes. Despite those conversations and multiple revisions of the MOU draft, what we discovered was ultimately different than the path our planning had indicated. Instead of creating an historical social network using the five decades of photographs we had prepared, we found that the history of the social network and the family and kinship relationships detailed through those images was deeply personal for the community living in the region today. We found out the hard way that those kinships reflected economic changes in status and power, realignments among families and their communities, and new patterns in the social fabric formed by the warp of personal relationships and the weft of local institutions (schools, hospitals, and local governance). Revealing those changes was not always something that our partners wanted us to do, and these were not patterns we had sought to discover: they are simply there, embedded in the images and the relations among images.

These social changes in local alignments—tied in complex ways to marriages and separations, legal conflicts and resolutions, changes in ownership of residential and commercial interests, and other material reflections of that social fabric—remain highly charged and, for those continuing to live in the area, they revealed potentially unexpected parts of the lived realities and values of the place. As a result, even though we had an MOU that worked for the technical details of the project, we could not find common ground for how to handle the competing social and ethical values of the project.

As we problem-solved, we tried to describe new forms of restriction and to generate appropriately sensitive guidelines to handle future use and access, but it turned out that all of these approaches were threatening to the values of a tightly knit community. They, rightly, want to tell their story, and so many people have told it so poorly for so long that they wish to have sole access to the materials from which the narratives are assembled. As researchers interested in open access and stable platform management, we have disagreements with the scholarly and archival implications of this decision, but we ultimately respect the resolve and underlying values that accompany the difficult choices PMSS makes about its public audiences and the corresponding goals it maintains for its collections. Interestingly, Wykle has come to view our work with PMSS collections as another form of the material and cultural extraction that has dominated the region

[14] See, for another example of the ethical quandaries that may be associated with legal applications of machine learning techniques, Ema et al. 2019.

for generations. While we see our work in light of preservation and access as well as our lasting commitment to PMSS and the region, we have also come to recognize the powerful explanatory force that the idea of "extraction" has become for the communities in a region that has suffered many forms of extraction industries' negative effects. In acknowledging the limitations of our own efforts, we would posit that our case study offers a counter-example to works that suggest how AI systems can be designed automatically to meet the needs of their constituents (Winfield et al. 2019). We tried to use a design approach to address our research goals and our partner's needs, and it turned out that the dynamically constructed and evolving nature of those needs outstripped the capacity we could build into our available system of machine learning.

The divergence of our goals has led the collaboration to an impasse. Given that we had already outlined further steps in our initial documents that could not be satisfied after the partners identified their divergent intentions, the collaborative scope the partners initially described was not completely fulfilled. The divergence of goals became stark: as researchers interested in the relevance and sustainability of these archives, we were moving the collections toward a more accessible and comprehensive platform with open documentation and protocols for future development. By contrast, the PMSS staff were moving toward more stringent and local controls over access to the archives in order to limit dissemination. At this juncture, we had some negotiating to do. First, we made the ContentDM instance a password protected and not publicly accessible (private) sandbox rather than a public instance of a virtual digital collection. As PMSS owns the material, they decided shortly thereafter to issue a take-down order of the ContentDM instance, and we complied. As the ContentDM materials were ultimately accessible in the public domain on their live site, this decision revealed how personal the challenges had become. Nothing included in the take-down order was unique or new material—rather, the ContentDM site simply provided a more accessible format for existing primary material on the WordPress site, stripped of its interpretive and secondary contexts.

If there is a silver lining, it lies in this context for use: the "academic divorce" we underwent by discontinuing our collaboration has made it possible for us to continue conducting research on the publicly available archival materials without being obligated to host a live and dynamic repository for further materials. As a result, we can test best-approaches without having to worry about pushing them to a live production site. Within this constraint, we aim to continue re-creating the historical social network without compromising our partners' needs for privacy and control of their production site. The mutual decision to terminate further partnership activities based in archival development arose because of these differing paths forward. That decision meant that any further enrichment of the archival materials would not become publicly available, which we saw as a penalty against using the archive at a moment when archives need as much advocacy and visible support as possible.

Under these constraints of private accessibility, we have continued to work on the AWS ReKognition pipeline and have successfully identified all of the faces of named people featured in the archive, with face and name labels now associated with over 1900 unique images. Our next step, delayed to Spring 2021 as a result of the COVID-19 pandemic, includes the creation of an associative network that first identifies unnamed faces in each image using unique identifiers. The second element of that process will be to generate an historical social network using the co-occurrence among those faces as well as the faces of named people in the available images. Given that our metadata enrichment has already included date associations for most of the images, we are confident that we will be able to reconstruct historically specific networks for a given year or range of years, and moreover, that the association between dates and named people will help us

to identify further members of the community who are not currently named in the photographs because of the small groups involved in activities and clubs, as well as the generally limited student and teacher populations during any given year.

We are now far more sensitive to how the local concerns of this community shape our research methods and outcomes. The longer-term hope, one it is not clear at all that we will be allowed to pursue, would be to use natural language processing tools on the archive's textual materials, particularly named entity recognition and word vectors, to search and match images where known names occur proximate to the names of unmatched faces. The present goal, however, remains to create a more replete and densely connected network of faces and the places they occupied when they were living in the gentle shadows of Pine Mountain. In order to abide by PMSS community wishes for privacy, we will be using anonymized aggregate results without identifying individuals in the photographs. While this method has the drawback of not being able to reveal the complexity of the historical relations at the granular level of individuals, it will allow us to report on the persistence or variation in network metrics, such as network density, centrality, path length, and betweenness measures, among others. In this way, we aim to be able to measure and report on the network and its changes over time without reporting on individuals. We arrived at an anonymizing method as a solution to the dissolved partnership by asking about the constraints of FERPA as well as by looking back at federal and commercial facial recognition practices. In each case, the dark side of these technological tools remains one associated with surveillance, and in the language of Eastern Kentucky, extraction. We mention this not only to be transparent about our recognition of these limitations, but also in the hopes of opening a new dialogue with our partners that might stem from generating interesting discoveries without compromising their sense of the local ownership of their archival materials. Nonetheless, in order to report on the most interesting aspects, the actual people and their local histories of place, the work to be done would remain more at a human level than at a technical one.

Conclusion

In conclusion, our project describes a success that remains imbricated with a shortcoming in machine learning. The machine learning tasks and algorithms our project implemented serve a mimetic function in the distilled picture of the community they reflect. By matching historical faces to names, the project embraces a form of digital surrogacy: we have aimed to produce a meta-historical account of the present institution's social and cultural function as a site of social networking and local knowledge transmission. As Robyn Caplan and danah boyd have recently suggested, the "bureaucratic functions" these algorithms promote can be understood by the ways in which they structure users' behaviors (2018, 3). We would like to supplement Caplan and boyd's insight regarding the potential coercions involved in how data structures implicitly shape their contents as well as their users' behaviors. Not only do algorithms promote a kind of bureaucracy, to ends that may be positive and negative, and sometimes both at once, but further, those same structures may reflect or shape public behaviors and interactions beyond a single platform.

As we move between digital and public spheres, our work similarly shifts its scope. The research that we intended to have positive community effects was instead read by that very same set of people as an attempt to displace a community from the center of its own history. In other words, the bureaucratic functions embedded in PMSS as an institution saw our new approach to their storytelling as an unwanted and external intervention. As their response suggests, the internal and extant structures for governing their community, its stories, and the people who tell them,

saw our contribution as an effort to co-opt their control. Where we thought we were offering new tools for capturing, discovering, and telling stories, they saw what Safiya Noble has recently characterized in a specifically racialized context as "algorithms of oppression" (2018). Here the oppression would be geographic, socio-economic, and cultural, rather than racial; nevertheless, the perception that one is being oppressed by systems set into place by agents working beyond one's own community remains a shared foundation in Noble's argument and in the unexpected reception of our project. As we move forward with our own project into unknown territories, in which our work-products may never see the light of day because of the value conflicts bound up in making archival objects public and accessible, we have found a real and lasting respect for the institutional dependencies and emplacements within which we all do our work. We hope to channel some of those functions of emplacement to create new forms of accountability and restraint that will allow us to move forward, but at least for now, we have found with our project one limitation of machine learning, and it is not the machine.

References

Ahmed, Manan, Maira E. Álvarez, Sylvia A. Fernández, Alex Gil, Rachel Hendery, Moacir P. de Sá Pereira, and Roopika Risam. 2018. "Torn Apart / Separados." Group for Experimental Methods in Humanistic Research. https://xpmethod.plaintext.in/torn-apart/volume/2/.

Bailey, Ronald. 2017. "The Noble, Misguided Plan to Turn Coal Miners Into Coders." *Reason,* November 25, 2017. https://reason.com/2017/11/25/the-noble-misguided-plan-to-tu/.

Calo, Ryan. 2017. "Artificial Intelligence Policy: A Primer and Roadmap." *University of California, Davis Law Review* 51:399-435.

Caplan, Robyn and danah boyd. 2018. "Isomorphism through algorithm: Institutional dependencies in the case of Facebook." *Big Data & Society* (January-June): 1-12. https://doi.org/10.1177/2053951718757253.

Cassidy, Frederic G. et al., eds. 1985-2012. *Dictionary of American Regional English.* Cambridge, MA: Belknap Press. https://www.daredictionary.com.

Ema, Arisa et. al. 2019. "Clarifying Privacy, Property, and Power: Case Study on Value Conflict Between Communities." *Proceedings of the IEEE* 107, no. 3 (March): 575-80. https://doi.org/10.1109/JPROC.2018.2837045.

Harkins, Anthony and Meredith McCarroll, eds. 2019. *Appalachian Reckoning: A Region Responds to* ***Hillbilly Elegy***. Morgantown, WV: West Virginia University Press.

Hochschild, Arlie. 2018. "The Coders of Kentucky." *The New York Times,* September 21, 2018. https://www.nytimes.com/2018/09/21/opinion/sunday/silicon-valley-tech.html.

Joh, Elizabeth. 2018. "Artificial Intelligence and Policing: First Questions." *Seattle University Law Review* 41 (4): 1139-44.

Latour, Bruno. 2007. *Reassembling the Social: An Introduction of Actor-Network Theory.* New York: Oxford University Press.

Levendowski, Amanda. 2018. "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem." *Washington Law Review* 93 (2): 579-630.

Mukurtu CMS. https://mukurtu.org/. Accessed December 12, 2019.

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

Passamaquoddy People. "Passamaquoddy Traditional Knowledge Labels." https://passamaquoddypeople.com/passamaquoddy-traditional-knowledge-labels Accessed December 12, 2019.

Risam, Roopika. 2015. "Beyond the Margins: Intersectionality and the Digital Humanities." *DHQ: Digital Humanities Quarterly* 9 (2). http://digitalhumanities.org/dhq/vol/9/2/000208/000208.html.

Robertson, Campbell. 2019. "They Were Promised Coding Jobs in Appalachia. Now They Say It Was a Fraud." *The New York Times*, May 12, 2019. https://www.nytimes.com/2019/05/12/us/mined-minds-west-virginia-coding.html.

Sabharwal, Anil. 2016. "Moving on from Picasa." Google Photos Blog. Last modified March 26, 2018. https://googlephotos.blogspot.com/2016/02/moving-on-from-picasa.html.

Sabharwal, Arjun. 2015. *Digital Curation in the Digital Humanities: Preserving and Promoting Archival and Special Collections.* Boston: Chandos.

Stephan, Karl D., Katina Michael, M.G. Michael, Laura Jacob, and Emily P. Anesta. 2012. "Social Implications of Technology: The Past, the Present, and the Future." *Proceedings of the IEEE 100, Special Centennial Issue* (May): 1752-1781. https://doi.org/10.1109/JPROC.2012.2189919.

United States Department of Justice. 2008. "Guidelines for a Memorandum of Understanding." https://www.justice.gov/sites/default/files/ovw/legacy/2008/10/21/sample-mou.pdf.

———. 2017. "Sample Memorandum of Understanding." http://www.doj.state.or.us/wp-content/uploads/2017/08/mou_sample_guidelines.pdf.

Vance, J.D. 2016. *Hillbilly Elegy: A Memoir of a Family and Culture in Crisis*. New York: Harper.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. New York: W.H. Freeman and Co.

Winfield, Alan F., Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. "Machine Ethics: the design and governance of ethical AI and autonomous systems." *Proceedings of the IEEE* 107, no. 3 (March): 509-17. https://doi.org/10.1109/JPROC.2019.2900622.