Audiovisual Algorithms

New Techniques for Digital Processing

by
Annie Schweikert

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Moving Image Archiving and Preservation Program
Department of Cinema Studies
New York University

6 May 2019

Advisor: Nicole Martin

Table of contents

Abstract

Born-digital audiovisual accessions have increasingly begun to outstrip analog accessions in archives, representing not only a change in format but a change in scale. In order to understand—not just preserve—the tsunami of born-digital content, archivists must take advantage of the format's possibilities for automated description and analytical work, through automation, artificial intelligence, and machine learning. This thesis grounds the reader in an understanding of recent technological innovations, surveys the tools and methods available to process digital audiovisual content; evaluates ethical concerns inherent in these approaches; and examines how archivists can effectively and responsibly put these tools into practice.

Acknowledgements

I am grateful for the following people, each deeply kind, generous, and wise:

- My thesis advisor, Nicole Martin
- My classmates Daniela Calle, Anne-Marie Desjardins, Christine Gennetti, Ari Greenberg, Jeff Lauber, Miles Levy, Sigga Regína Sigurþórsdóttir, and Draye Wilson
- My professional supervisors past and present, in particular Dave Rice and Pam Wintle
- Connor Hoge
- Coffee

I also give my thanks to everyone else who offered their advice and support for this project, in particular those who gave their time to speak with me one-on-one. Please see the bibliography for a

complete list.

Introduction

Born-digital audiovisual accessions have increasingly begun to outstrip analog accessions in archives. This shift has spurred archivists to adapt workflows and archival concepts to digital materials. But born-digital material represents not only a change in format—it is a change in scale. Cheap to produce and cheap to store, digital counterparts to analog film and video run so long and large that traditional processing methods (such as finding aids and catalogs) cannot scale to meet the demands of the material. The ability to process unstructured data has huge potential for large-scale digital collections as well as for collections of complex media, such as audiovisual materials. Automations and machine learning techniques are likely the next frontier for rendering these otherwise-opaque materials discoverable.

In order to understand—not just preserve—the tsunami of born-digital content, archivists must take advantage of the format's possibilities for automated analytical work in the form of machine learning, natural language processing, facial and object recognition, closed captioning extraction, and automated transcription, among other methods. Stakeholders in this goal include archivists, librarians, digital humanists, and machine learning researchers.

In this thesis, I describe the parameters of automation and machine learning; identify specific ways in which humans can automate digital audiovisual content description discovery, through textual and audiovisual methods; provide frameworks for auditing and integrating machine learning into archival workflows; and address ethical concerns pertaining to archives and libraries.

I. Definitions and context

Automating and integrating basic cataloging, acquisition, and access tasks is a major part of modern archives, library, and museum work. Artificial intelligence and machine learning tools are considered by many to be the next step in streamlining workflows and easing workloads. However, artificial intelligence applications remain largely theoretical rather than practical for most institutions, even those that have embraced automation and other processing strategies that take advantage of digital materials.

1. Automation

Automation, or the use of machines to perform tasks that would otherwise fall to humans,[1] began to be widely implemented in libraries and archives with the advent of computer technologies in the 1960s. One of the earliest and most famous examples of library automation is the MARC (MAchine-Readable Cataloging) format, developed in 1968 by Henriette Avram at the Library of Congress. MARC compiles bibliographic records into a centralized, searchable database. In doing so, it not only replaced the manual system of card cataloging but created links between previously-isolated keywords.[2] Avram not only saved librarians time on daily tasks, but built the foundation of a collection of tagged and structured data that could be shared by libraries across the country. Her system spurred automation and connectivity of more and more daily tasks over the next several decades, enhanced by evolutions in computer hardware and software.

[1] Mikell P. Groover, "Automation," *Britannica.com,* Encyclopedia Britannica, 22 Mar. 2019, https://www.britannica.com/technology/automation
[2] Matt Schudel, "Henriette Avram, 'Mother of MARC,' Dies," *Information Bulletin,* Library of Congress, May 2006. Reprinted from *The Washington Post,* page B06, 28 Apr. 2006. https://www.loc.gov/loc/lcib/0605/avram.html

The advent of the Internet provided a major field on which to share data and reduce local workloads, and spurred the development of many data exchange frameworks in the vein of MARC. However, the Internet also represented a model in which closed databases, such as library catalogs, were disadvantaged against results that could be retrieved by indexers such as Google. The public was quicker than institutions to adjust, though libraries in particular are beginning to implement networked data models such as Resource Description Framework (RDF) and International Image Interoperability Framework (IIIF).[3]

On a practical level, present-day automation in libraries and archives typically takes the form of consolidated, all-purpose products that incorporate a set of metadata standards. In libraries, individual automated tasks have been collated and synchronized into platforms known as Integrated Library Systems; these systems are the de facto standard for libraries seeking to manage their cataloging, circulation, and search functions.[4] In archives and museums, Collections Management Systems[5] and Digital Asset Management Systems[6] have proliferated to perform several different combinations of tasks, including accessions, loans, ingest, access, and long-term preservation. These systems allow institutions to track individual objects from acquisition to deaccessioning, all the while tagged and discoverable with structured metadata.

While many automation systems operate as black boxes, other systems reveal how small automation tasks can be linked into a system more powerful than the sum of its parts. Archivematica, an open-source digital preservation system that creates structured packages that store digital objects,

---

[3] Catherine Nicole Coleman, "Library-Inspired Artificial Intelligence: Discovery, Part 1," *Digital Library Blog*, Stanford University Libraries, 22 Oct. 2018, http://library.stanford.edu/blogs/digital-library-blog/2018/10/library-inspired-artificial-intelligence-discovery-part-1

[4] Marshall Breeding, "Library Systems Report 2018," *American Libraries Magazine,* 1 May 2018, https://americanlibrariesmagazine.org/2018/05/01/library-systems-report-2018/

[5] Ashley Blewer, "The Collection Management System Collection" (collaborative spreadsheet), created Aug. 2017, https://docs.google.com/spreadsheets/d/1cXOug3qM0pNNeD_wssiVEv9c0W1Y5I1VDTnSPTk7fb4/edit#gid=0

[6] Leala Abbott, "The DAM List" (spreadsheet), created Feb. 2011, https://docs.google.com/spreadsheets/d/1xRwkQVluqtlLVeuLqHtx3EtZeNAye3n_7BwR13GKwm0/edit#gid=0

operates by breaking down its larger function into "granular system tasks," or microservices. To create a package, Archivematica executes a long list of discrete microservices in one-task units. For example, in one sequence of steps, the program scans for viruses (one microservice), cleans up filenames (another), and identifies file formats (a third).[7] Once tasks are broken down into small units and automated, they can be combined, reordered, and swapped out in infinite combinations, suitable for a range of collections and institutions.

Whatever the exact tool or framework, the appeal of automation is obvious. By automating work that used to be manual, archivists and librarians can spend less time on mechanical tasks and can expect more uniform results. In addition, the structured data created or required by automation generates an exponential number of uses. MARC was the first tool that required structured, machine-readable data. This structured data could be then packaged and exchanged between different institutions, reducing redundant workloads (if one library has cataloged a book, the other library can copy that catalog record) while also building information connections between institutions (if a patron at one library needs the second edition of a book, a librarian can find that edition at another library and arrange an interlibrary loan).

2. Artificial intelligence

Though it is often referred to as a next step beyond automated systems, artificial intelligence is not a natural outgrowth of automation so much as a complementary concept. As a blanket term for several narrower fields of study, and long since entered in the public lexicon, "artificial intelligence" tends to shift in meaning depending on who is using the label. This paper will use a textbook

[7] "Micro-services," *Archivematica Development Wiki*, last modified 14 Aug. 2015, https://wiki.archivematica.org/Micro-services

definition of artificial intelligence, as computing that creates and trains an "intelligent agent," that is, a machine that receives signals from its environment and uses those signals to take actions.[8] This intelligence differs from rote programming in that it is able to interpret signals correctly and act rationally outside of the bounds of strict computer logic.[9]

AI conceptually dates back to the 1940s—near the beginning of the modern computing movement—with early and influential contributions made by Alan Turing towards the end of his life. AI encompasses several disciplines, such as the six defined in one major textbook as natural language processing, or linguistic analysis; knowledge representation, or storage of knowledge; automated reasoning, or the ability to answer questions and draw conclusions; machine learning, to detect patterns; computer vision, to "see" images and video; and robotics, to "move."[10] Many of these subfields have applications within libraries and archives. For example, advancements of computer vision have facilitated the development of optical character recognition, which has been employed to great effect with books and typed manuscripts. As textual documents are extremely consistent—black text on white backgrounds, with uniform scale, orientation, layout, and typeface or handwriting style—it is relatively simple to train a computer to "see" an image with text and translate it into a structured textual representation.[11] Other applications involve parsing text for meaning, such as using natural language processing to extract people, organizations, topics, and keywords from large corpuses of material. Such tools have become so commonplace as to be bundled and offered within open-source software packages, as was the goal of the Mellon-funded BitCurator NLP project.[12]

---

[8] Stuart J. Russell and Peter Norvig, *Artificial intelligence: a modern approach*, 3rd ed. (Upper Saddle River, New Jersey: Pearson Education Limited, 2010), viii.
[9] Russell and Norvig, 5.
[10] Russell and Norvig, 2-3.
[11] Lienhart, Rainer W., and Frank Stuber. "Automatic text recognition in digital videos." In Image and Video Processing IV, vol. 2666, pp. 180-189. International Society for Optics and Photonics, 1996, 180-182, 186.
[12] "BitCurator NLP," BitCurator.net, BitCurator, accessed 2 May 2019, https://bitcurator.net/bitcurator-nlp/

The history of artificial intelligence lacks a complete academic accounting, which this paper does not have the scope to correct.[13] However, one essential quality of American AI is its link to corporate and military development, particularly in the service of defense. The first artificial intelligence program was developed by employees of the RAND Corporation and funded by the Air Force.[14] The Department of Defense has long funded AI research through its Defense Advanced Research Projects Agency (DARPA), including an early version of what became Apple's Siri. DARPA has attracted talent by positioning itself as a source of funding without the corporate strings of tight deadlines and profitability; in 2018, DARPA announced it would fund two billion dollars' worth of projects over five years, apparently to compete with China.[15] AI has been named a national priority by the Obama and Trump presidential administrations.[16][17] In 2019, the International Data Corporation estimated that 7.8% of worldwide AI funding went towards "automated threat intelligence and prevention systems" (second only to customer service bots); 60% of this funding came from the United States.[18]

Another essential quality of the American AI landscape is its dominance by tech giants: Amazon, Apple, Facebook, Google, IBM, Microsoft, and, more recently, Tesla and Uber. AI-specific budgets at these companies are hard to estimate, but other actions make clear the depth of these investments. Big tech is famous for intense recruiting strategies and is known to "plunder" academic

---

[13] AI is usually presented as a series of technological advancements without context, and ignoring non-Western and Anglo contributions. Mentions of non-Western AI projects often posit a competition between the United States and China, whose government in 2017 outlined a "three-step roadmap" to economic transformation through AI. (Arjun Kharpal, "China wants to be a $150 billion world leader in AI in less than 15 years," *Tech Transformers*, CNBC, 21 Jul. 2017, https://www.cnbc.com/2017/07/21/china-ai-world-leader-by-2030.html)

[14] Jonnie Penn, "AI thinks like a corporation—and that's worrying," *Open Future* (blog), *The Economist,* 26 Nov. 2018, https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying

[15] This announcement was in the face of the withdrawal of some key big tech companies from government-funded projects. (Drew Harwell, "Defense Department pledges billions toward artificial intelligence research," *The Switch* (blog), *The Washington Post,* 7 Sept. 2018, https://www.washingtonpost.com/technology/2018/09/07/defense-department-pledges-billions-toward-artificial-intelligence-research/)

[16] Ajay Agrawal, Joshua Gans, and Avi Goldfarb, "The Obama Administration's Roadmap for AI Policy," *Harvard Business Review,* 21 Dec. 2016, https://hbr.org/2016/12/the-obama-administrations-roadmap-for-ai-policy

[17] Cade Metz, "Trump Signs Executive Order Promoting Artificial Intelligence," *The New York Times,* 11 Feb. 2019, https://www.nytimes.com/2019/02/11/business/ai-artificial-intelligence-trump.html

[18] "Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $77.6 Billion in 2022, According to New IDC Spending Guide," *IDC Corporate USA,* 19 Sept. 2018, https://www.idc.com/getdoc.jsp?containerId=prUS44291818

departments and conferences for employees.[19] Though potential competitors have proliferated in recent years—in the form of smaller but well-funded AI startups[20]—these companies are likely to be quickly snapped up by tech giants. Amazon, Apple, Facebook, Google, IBM, and Microsoft acquired a combined 40 AI companies from 2012 through 2017.[21] Most artificial intelligence tools available for general use have been developed by these companies, for reasons including the fact that big tech has access to huge amounts of personal data through their own users and services. Though these companies' research and commercial divisions are technically separate, this separation is muddied by the fact that the parent company owns all intellectual property developed in both.[22]

Some subfields of AI are more relevant to libraries and archives than others. For example, natural language processing provides a linguistic entry point to large amounts of data.[23] On the other hand, automated reasoning, which can be used to power chatbots or other artificial interactions, is antithetical to current conceptions of the human reference interview. In addition, the development of tools for libraries, archives, and museums currently relies largely on grant funding, volunteer work, and collaboration with adjacent fields. Investing in AI for libraries and archives does not offer the same financial return as the same investment in retail, big tech, and the military. These arenas of funding, research, and development have shaped the direction, creation, and dissemination of artificial intelligence and machine learning tools. It is important to consider this influence over the qualities of these tools as we examine them for adoption within libraries and archives.

---

[19] "Google leads in the race to dominate artificial intelligence," *The Economist* (print), 7 Dec. 2017, https://www.economist.com/business/2017/12/07/google-leads-in-the-race-to-dominate-artificial-intelligence

[20] Jean Baptiste Su, "Venture Capital Funding For Artificial Intelligence Startups Hit Record High In 2018," *Forbes.com,* 12 Feb. 2019, https://www.forbes.com/sites/jeanbaptiste/2019/02/12/venture-capital-funding-for-artificial-intelligence-startups-hit-record-high-in-2018/#19e6357f41f7

[21] "Big Tech In AI: What Amazon, Apple, Google, GE, And Others Are Working On," *CB Insights,* 12 Oct. 2017, https://www.cbinsights.com/research/top-tech-companies-artificial-intelligence-expert-intelligence/

[22] Olivia Solon, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent," *NBC News,* 12 Mar. 2019, https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[23] Catherine Nicole Coleman, "Artificial intelligence and the library of the future, revisited," *Digital Library Blog*, Stanford University Libraries, 3 Nov. 2017, http://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited

3. Machine learning

The theory of machine learning, sometimes known as data mining or predictive analytics, dates to the early decades of AI.[24] Machine learning differs from generic artificial intelligence, as well as simpler automation solutions, because its process of improvement is driven by the machine (or program, or software) itself. Automated solutions are hard-coded, step by step, by human programmers. Such machines require the human programmer to anticipate every step of every solution for each situation possible, both for today and for the future. In cases where input is predictable and output is standardized, this approach is adequate—for example, in an archival environment that normalizes all the digital video files it receives to one codec/container combination. In other cases, the cost of tinkering with a program to account for new developments and outliers is prohibitive. A piece of software that can learn becomes more accurate when it is fed examples and data, reducing the burden of design.[25]

Machine learning is also popular for tackling tasks that are too complex for humans to model step by step.[26] For example, say someone wanted to filter a collection of family photos down to just those that featured her grandmother. Could we even break down the multitude of intuitive, instantaneous steps it takes for us to recognize a family member? What about recognizing the same grandmother as she aged, or was just photographed from different angles? It is impossible to represent the full cognitive process of human recognition. It would be far easier to feed several pictures of the same person to a program and ask the program to think like a human. This solution is possible—within limits—with machine learning.

---

[24] Russell and Norvig, 2-3.
[25] Russell and Norvig, 693.
[26] Russell and Norvig, 693.

This shift in method that machine learning represents—from painstakingly refining algorithms, to feeding the same algorithms massive amounts of data and asking the algorithms to refine themselves—only became possible with the availability of extremely large sets of training data, or "big data." Big data, along with sharp growths in computing power, have been suggested responsible for the surge in machine learning projects from the early 2000s through the present.[27] This growth is not new in the history of artificial intelligence; research enthusiasm for AI has gone through several cycles of heightened and depressed funding over the next half-century, including an "AI Winter" in the 1980s after a wave of industry-funded projects failed to deliver financial results.[28]

However, this cycle of AI is the only one that has incorporated machine learning into everyday commercial products. Familiar applications of machine learning include security features, such as email spam filters and facial recognition that can unlock phones, as well as commercial applications such as playlists that recommend new music based on your listening habits and chatbots that attempt to provide online customer service. Machine learning can also dramatically improve tools already in use,such as OCR; while it would be near-impossible to program a set of rules for recognizing handwritten text, machine learning models can be trained to interpret cursive with a set of already-transcribed examples.

Despite these integrations and successes, the actual state of machine learning is less advanced than its potential suggests. Even corporate interests are not enough to ensure high performance; despite nearly endless money and research time at their disposal, Uber and Tesla have killed people in the process of developing a self-driving car, and the technology remains elusive at best.[29] AI-related

---

[27] Charlie Harper, "Machine Learning and the Library or: How I Learned to Stop Worrying and Love My Robot Overlords," *Code4Lib Journal* 41, (Aug. 2018). https://journal.code4lib.org/articles/13671

[28] Russell and Norvig, 16-24.

[29] Timothy B. Lee, "The hype around driverless cars came crashing down in 2018," *Ars Technica,* 30 Dec. 2018, https://arstechnica.com/cars/2018/12/uber-tesla-and-waymo-all-struggled-with-self-driving-in-2018/

concepts are perennial entries in the "Hype Cycle" graphs produced by the technology research firm Gartner; machine learning remains in stages such as "inflated expectations" and "trough of disillusionment," having not yet landed in the "plateau of productivity."[30]

4. How does machine learning work?

Because the mechanics of machine learning have such an impact on their usefulness and adoption within libraries and archives, we will pause for an overview of the way machine learning works.

A. Algorithms

The set of steps a machine takes to process the input it receives and turn it into output is called an algorithm. An algorithm can be as simple as basic math—adding two numbers together (input) to yield a third (output) is an algorithm. Algorithms can be fixed, as in traditional approaches to computing, or they can evolve, as in the training process of machine learning. The term "algorithm" formally refers to an actual mathematical construct, but it is commonly used to imply the implementation of the algorithm within a specific context and the training—or unique evolution—it undergoes for that application.[31]

In the training process, a programmer starts with an initial algorithm, feeds it data, and uses that data to produce a new, more refined algorithm. Algorithms that have gone through training are

[30] Kasey Panetta, "5 Trends Emerge in the Gartner Hype Cycle for Emerging Technologies, 2018," *Gartner, Inc.,* 16 Aug. 2018, https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/

[31] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society* 3, no. 2 (2016), 2-3.

also referred to as models, implying a more specific application of the algorithm to the data training set.[32] Training means that algorithms will optimize themselves to the data they are provided. However, implementing machine learning usually does not mean starting from scratch. Thousands of published algorithms represent different methods of generating useful information, evaluating that information, and doing it the most efficient way possible.[33]

Though each model resulting from the machine learning training process is different, all models begin with an underlying algorithm. There are many algorithms, each of which fits into one of three training methods: supervised learning, reinforcement learning, and unsupervised learning.

In supervised learning, the program is provided with data which it attempts to classify, and is corrected after every attempt. Imagine trying to train a machine to pick out all the photos of your grandmother from a collection of family photos. A supervised learning method would require that each photo be tagged as containing the grandmother (or not). Using this information, the machine would be able to receive immediate feedback and correct its algorithm slightly for each success or failure. This type of supervised learning algorithm is called "classification."[34]

Reinforcement learning provides less explicit correction, in the form of rewards or punishments, and asks the machine to figure out what part of its attempt was correct or incorrect. In reinforcement learning, a machine might attempt to sort photos by whether or not they contain the grandmother's face; at the end, the results might be graded on a pass/fail basis, without individual correction.[35] However, reinforcement learning is most common in robotics situations—for example, asking a robot to solve a maze and grading it on whether it succeeded.[36]

---

[32] Harper.
[33] Domingos, 78-79.
[34] Russell and Norvig, 694-695.
[35] Russell and Norvig, 694-695.
[36] "How to choose algorithms for Azure Machine Learning Studio," *Microsoft Azure Machine Learning Studio Documentation,* 3 Mar. 2019, https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

Unsupervised learning supplies no explicit feedback and asks the machine to group potentially related input. This approach might result in the machine successfully sorting different examples of the grandmother's face together but not being able to label them as the grandmother. Unsupervised learning algorithms are known as "clustering" methods.[37]

Selecting an algorithm often depends on what the data itself looks like. Supervised learning in particular requires that training data be extensively tagged and weeded so that the machine can check its own work. Reinforcement learning is best suited to environments where the machine reaches a binary conclusion after a series of decisions, such as getting through a maze or winning a game. Unsupervised learning, or asking the machine to recognize clusters on its own, is the most flexible method of training but can have the most variable results. In some cases, the machine may recognize patterns that a human would otherwise not see, such as in a medical application that used machine learning to recognize diseased retinas and, as a byproduct, found that age, gender, and BMI correlated to retinal problems.[38] In other cases, the machine may come up with technically correct but functionally useless classifications. Even when clustering appears to work, human intervention will always be required to make sense of the results and to avoid assumptions about causation.

Selecting an algorithm does have implications for the final model. Different algorithms have different levels of accuracy, which can either make the final model itself more accurate or—depending on the problem it is meant to address—overfit data to a specific scenario. In other cases, the choice of algorithm will require varying amounts of training time. Though more training time usually yields a more accurate model, there are cases in which long training times combined with a large dataset are

---

[37] Russell and Norvig, 694-695.

[38] Ryan Poplin et al, "Predicting cardiovascular risk factors from retinal fundus photographs using deep learning," arXiv preprint arXiv:1708.09843 (2017), cited in Coleman, "Library-Inspired Artificial Intelligence."

impractical for use. Algorithms also have different numbers of parameters, which are the settings that can be adjusted at the start of a training session.[39]

B. Data

Huge datasets, containing billions and trillions of words, images, and other content, have increasingly become available in the last two decades. The impact of this amount of data appears to outweigh any other tinkering that can be done with algorithms; even unlabeled data in large amounts can turn a "mediocre" algorithm into one that outperforms the "best known" version.[40] Time and memory are commonly known as the de facto limitations of advancements in computer science; data has become as indispensable to machine learning, and the lack of it just as limiting.[41] Datasets are so important to the training process that companies often lock them down much more rigorously than any other component of a machine learning model.[42]

Training data must be collected, cleaned, and sometimes labeled before use—a process far more time-consuming than the actual time spent training models. Cleaning data is the process of removing misleading information and noise from the data, from spelling mistakes to incorrect labeling to audio or video so low-quality as to be useless. For example, to train a model to isolate a single person's speech from a mixture of sounds, engineers at Google had to first find videos of lectures and talks; eliminate any videos that were of low visual quality; extract segments with only one speaker in the frame; and ensure that these segments were only of one audible voice, with no background noise.

---

[39] "How to choose algorithms for Azure Machine Learning Studio."
[40] Russell and Norvig, 27-28.
[41] Domingos, 85.
[42] Max Grigorev, "Keeping up with AI in 2019," *The Launchpad* (blog), 14 Feb. 2019, https://medium.com/thelaunchpad/what-is-the-next-big-thing-in-ai-and-ml-904a3f3345ef

Only then, after the team had cleaned and cut down 100,000 videos to 2000 hours of video clips, could training of the model begin.[43]

Labeling data is another large time investment. Data labeling is typically repetitive and rote, and includes tasks such as labeling every object in a video or, for better data quality, approving or correcting other people's first-round labels.[44] Some commercial artificial intelligence companies skirt this investment by hiring workers for extremely low wages through Mechanical Turk,[45] factories,[46] and even prisons.[47] Other research projects use open, high-quality datasets such as the Stanford Vision Lab's ImageNet image dataset (itself largely labeled by Mechanical Turk workers)[48] or the TRECVID video dataset,[49] both of which were made available for competitions. However, the right dataset does not exist for every question—or it does exist, but it is locked down in the interest of competition by whichever company made it—meaning that many researchers must create their own, or work with unsupervised (unlabeled) learning methods.

Though big data allows for sophisticated machine learning applications, the effectiveness of a dataset does not necessarily increase in proportion with its size. As Pedro Domingos explains, the complexity of sophisticated algorithms require much more data to train a useful model. In theory, therefore, more training data should allow researchers to start with already-sophisticated algorithms and feed it huge amounts of data, to yield a complex model. At a certain point, however, the amount

---

[43] Inbar Mosseri and Oran Lang, "Looking to Listen: Audio-Visual Speech Separation," *Google AI Blog,* Google, 11 Apr. 2018, https://ai.googleblog.com/2018/04/looking-to-listen-audio-visual-speech.html
[44] Tom Simonite, "To Make AI Smarter, Humans Perform Oddball Low-Paid Tasks," *Wired,* 9 Feb. 2018, https://www.wired.com/story/behind-artificial-intelligence-lurk-oddball-low-paid-tasks/
[45] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. "A data-driven analysis of workers' earnings on Amazon Mechanical Turk," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, 1.
[46] Li Yuan, "How Cheap Labor Drives China's A.I. Ambitions," *The New York Times,* 25 Nov. 2018, https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html
[47] Angela Chen, "Inmates in Finland are training AI as part of prison labor," *The Verge,* 28 Mar. 2019, https://www.theverge.com/2019/3/28/18285572/prison-labor-finland-artificial-intelligence-data-tagging-vainu
[48] John Markoff, "Seeking a Better Way to Find Web Images," *The New York Times,* 19 Nov. 2012, https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html
[49] "TRECVID data availability by year and task," *TRECVID,* National Institute of Standards and Technology, 17 Sept. 2018, https://trecvid.nist.gov/past.data.table.html

of *time* it takes to train that already-sophisticated algorithm cuts into its gains. For this reason, researchers start with the simplest algorithms and only move onto more advanced starting points if the results are unsatisfactory. Solving the processing time bottleneck will allow researchers to start with more sensitive and adjustable algorithms, and presumably yield much more sensitive results.[50]

C. Neural networks and deep learning

A few techniques exist to combine and package multiple algorithms for more complex results. One of these is the artificial neural network, which is often described as a structure that attempts to model human thought. Neural networks are trained by a series of algorithms to model the human brain's neurons (which receive, process, and produce data) and the links between these neurons (which process data). In a neural network, an artificial neuron (also called a node) acts as a carrier for the data. The data is transformed through multiple connected layers of neurons. The first layer of neurons represents the input data; the last layer of neurons represents the output data; and "hidden" layers in the middle represent chances for the neurons to process and transform the data.[51] "Deep learning" refers to neural networks with multiple layers, and is largely used as a marketing term. Neural network and deep learning models allows for more subtle and complex processing, but also require far more processing time.[52]

Artificial neural networks enable unsupervised, inference-based learning for more abstract concepts and at higher levels than non-neural network methods.[53] They are particularly well-suited to computer vision, speech and audio recognition, and natural language processing—that is, areas of

---

[50] Domingos, 81-85.
[51] 3Blue1Brown, "But what *is* a Neural Network? | Deep learning, chapter 1," YouTube video, 19:13, *3Blue1Brown Series* season 3, episode 1, 5 Oct. 2017, https://www.youtube.com/watch?v=aircAruvnKk&feature=youtu.be
[52] Robert D. Hof, "Deep Learning," *MIT Technology Review,* 23 Apt. 2013, https://www.technologyreview.com/s/513696/deep-learning/
[53] 3Blue1Brown.

great concern to audiovisual materials.[54] They also tolerate noisy data with better resilience than simpler algorithmic learning models.[55]

[54] Weijiang Feng, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo, "Audio visual speech recognition with multimodal recurrent neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2017, 681.

[55] Russell and Norvig, 728.

II. Applications for audiovisual archives

1. Tools for description

Audiovisual content processing and description present several opportunities for description by automation and machine learning. Some methods are entirely based on the image and sound content, such as facial recognition and audio indexing; other methods rely on text, whether inherent in the digital file or extracted from the audio or video.

A. Text-based methods

Text processing methods extract existing text from unindexed parts of the video file, such as the closed caption lines in a television signal or the BEXT chunk in a Broadcast WAVE file. These methods may yield unstructured (e.g. free text) or structured (e.g. XML-formatted) text. Text-based content extraction methods are often the simplest option for those looking to automate content description, as they take advantage of metadata and transcripts, and suitable for integration into existing workflows, as many content management systems rely upon text-based search and standardized metadata fields for access.

Many of these methods rely on natural language processing (NLP), a form of artificial intelligence that attempts to understand human language. This task is complex, given that the way humans speak and write is characterized by a high degree of ambiguity and shifting meanings, as opposed to the formal programming languages that typically shape human interactions with machines.[56] The ability to extract meaning from text-based information relies upon a range of abilities, including

[56] Russell and Norvig, 860-861.

recognizing different languages, categorizing the genre of the writing, and performing sentiment analysis (classifying sections of text as generally positive or negative). Natural language processing also encompasses information retrieval, or correctly interpreting a user's query and fetching relevant documents.[57] Given the complexity of human language, NLP usually does not rely on hard-coded rules, but probabilistic models such as hidden Markov models.[58] NLP tools do not necessarily incorporate machine learning processes, though the two—both subfields of AI—may be combined.

I. Metadata extraction

Metadata extraction, or reporting values for metadata fields embedded in digital files, can be a simple yet powerful tool for both born-digital and digitized objects. All digital objects travel with metadata fields about the technical characteristics and creation of the file, such as creation date. In the case of born-digital files, this information describes the original object, while for digitized files it describes the surrogate. Born-digital and digitized files may also travel with descriptive and administrative data. While these metadata are not inherent to the digital object, they have often been added by creators or caretakers at different parts of the material's life cycle. Digital objects with embedded descriptive metadata are at an advantage for access and preservation, as they are fully self-descriptive and do not rely on linked databases or catalog records. In other cases, objects are described in a catalog record or other linked resource. The description itself requires human labor, but where labor has already been expended, it yields rich dividends for searching and summarizing.

Technical metadata can be parsed by a wide variety of tools, including the open source tools MediaInfo[59] and the File Information Tool Set.[60] Embedded descriptive metadata may also be parsed

[57] Russell and Norvig, 860-867.
[58] Russell and Norvig, 876.
[59] "MediaInfo," *MediaArea.net,* 2019, https://mediaarea.net/en/MediaInfo

by similar tools. In cases where information exists only in a catalog record or database, a variety of methods exist for extracting the metadata, each tailored to the specific database in use. Once extracted, metadata fields and values can be transformed into structured or textual formats, passed to databases for access and storage, and searched or sorted.

Technical and descriptive metadata provide multiple avenues of applications. Descriptive and administrative metadata is the most useful, often providing key information such as subject headings and provenance information. Though technical metadata is less explicit, it can be used to identify footage within certain parameters. For example, in environments with predictable equipment—such as a television station with in-studio footage and footage shot on location—a file's technical metadata can sort footage by location based on the type of camera reported by the file's technical metadata.[61] In other cases, administrative metadata holds clues to the content of the files, and can be extracted. For example, researchers at the NEH-funded Photogrammar project hosted by Yale found that cryptic-seeming location codes actually were clues for a historic classification system built in the 1940s.[62]

II. Metadata grooming

In cases where metadata already exists but is inconsistent, outdated, or otherwise problematic, metadata "grooming" or "cleaning" can be a proactive and programmatic way to address erratic or offensive description. Descriptive metadata issues appear across a spectrum of causes. One category of cases have to do with outdated conventions within the metadata. For example, women may be identified as "Mrs. [Husband's Firstname Lastname]" instead of by their own names. Or the data may

[60] "File Information Tool Set (FITS)," *Projects at Harvard,* Harvard University, 2019, https://projects.iq.harvard.edu/fits/home
[61] Conversation with Dave Rice, Oct. 2019.
[62] Conversation with Lauren Tilton and Taylor Arnold, 16 Apr. 2019.

be described according to conventions that are still in place, though many aspects of the conventions themselves are outdated; take the Library of Congress Subject Headings, which remain widely used despite the use of obsolete and offensive terminology to describe many groups of people. Projects such as Noah Geraci's "mrs" and "indigenous-lc-wikidata" model ways to correct these outdated metadata practices.[63] Other issues of outdated conventions may refer to local guidelines that have changed over time or been inconsistently applied, such as how long image titles should be or what uniform tags should look like; the Carnegie Museum of Art has worked to standardize these and other tags by scripting in a project using the Charles "Teenie" Harris archive.[64] Another category of metadata inconsistencies are caused by technological limitations. For example, incompatible character encodings between operating systems can lead to diacritics in non-English languages being flagged as "illegal characters." Such limitations reinforce a status quo where English reigns; addressing them programmatically is a way to treat materials with respect for their languages and origins.[65]

III. Transcript extraction

Closed captions, an accessibility requirement in both analog and digital TV broadcast standards, can serve as a ready-made transcript if extracted from the signal.[66] Closed captions can be extracted from analog signals by capturing line 21 of the broadcast signal during the digitization process,[67] or (for analog signals encoded for digital broadcast) decoding the user-data field of the

[63] Noah Geraci, "Programmatic approaches to bias in descriptive metadata" (presentation, Code4Lib 2019, San José, CA, Feb. 2019), https://osf.io/9uehx/
[64] "Teenie Week of Play" (GitHub repository), Carnegie Museum of Art, last updated Jan. 2019, https://github.com/cmoa/teenie-week-of-play
[65] Elvia Arroyo-Ramirez, "Invisible Defaults and Perceived Limitations: Processing the Juan Gelman Files," 30 Oct. 2016, https://medium.com/on-archivy/invisible-defaults-and-perceived-limitations-processing-the-juan-gelman-files-4187fdd36759
[66] Lalitha Agnihotri, Kavitha Vallari Devara, Thomas McGee, and Nevenka Dimitrova, "Summarization of video programs based on closed captions," in *Storage and Retrieval for Media Databases 2*001, vol. 4315, International Society for Optics and Photonics, 2001, 601.
[67] Doug Keltz, "Understanding & Troubleshooting Closed Captions" (presentation), July 2014, https://www.smpte.org/sites/default/files/section-files/2014_July_Closed_Captioning.pptx

MPEG-2 program stream.[68] Born-digital broadcast signals also carry the captions in the MPEG-2 program streams, and can be demuxed and decoded into textual transcripts.[69] This process can be accomplished with proprietary tools as well as open source tools such as FFmpeg.[70] Once extracted, these captions can, with some basic manipulations, be integrated into full-text search platforms and found by keyword search.

IV. Translation

Natural language processing through machine learning is also responsible for advances in machine translation in recent years. Prior to the adoption of machine learning techniques, attempts to translate using machines were usually limited to programs written step by laborious step according to the rules of grammar. With machine learning—enabled by the wide availability of text online in multiple languages—subtleties of syntax, grammar, and vocabulary are now much more apt to be translated correctly, because the tools can learn from correct translations. Machine translation can help make collections available to users who do not speak its language. However, machine translation will not preserve subtleties about the materials, such as intent, that remain the domain of human translators and speakers. In addition, machine translation is far better for "high-resource languages," or languages that have large concentrations of both resources and data, such as English, Spanish, and Chinese; translation remains subpar for "low-resource languages" such as Indonesian and Swahili.[71]

---

[68] SMPTE 334M, as described by Sarkis Abrahamian in "EIA-608 and EIA-708 Closed Captioning," Evertz, n.d., accessed 14 Oct. 2018, https://evertz.com/resources/eia_608_708_cc.pdf
[69] Agnihotri et al, 601.
[70] "Can ffmpeg extract closed caption data" (Stack Overflow question), posted by spinon, 3 Jul. 2010, https://stackoverflow.com/questions/3169910/can-ffmpeg-extract-closed-caption-data
[71] Julia Hirschberg and Christopher D. Manning, "Advances in natural language processing," *Science* 349, no. 6245 (2015), 261.

B. Audiovisual content-based methods

Audio, visual, and audiovisual processing methods translate pure image and audio content into descriptive metadata. These methods may yield content in the form of tags for faces and objects, or audio and video divided into logical segments. These methods differ from text-based methods, which simply find and index existing textual content, in that they generate new metadata from the audiovisual content of the file.

Audiovisual processing with machine learning can yield results for high-level or low-level features. High-level features, sometimes referred to as "semantic concepts," represent complex and complete concepts that humans express using natural language. Semantic concepts are how humans expect to search for and describe materials. Unfortunately, they are far more difficult for machines to identify than the low-level features that make them up, such as color, texture, and shape (for images) or phonemes (for words). This gap—between the basic, machine-identifiable characteristics of a video and the high-level human concepts the video represents—is known as the "semantic gap."[72]

Bridging this gap requires labeled data. For example, in order to recognize a specific voice (say, Terry Gross) in an audio recording, a machine must first be trained on examples of audio that have been identified as Terry Gross's voice. Unfortunately, audio rarely comes tagged in discrete, isolated units appropriate for this kind of training, and humans often must take the pre-training step of labeling audio. Programs such as the Audio Tagging Toolkit, funded by the High Performance Sound Technologies for Access and Scholarship (HiPSTAS) project (led by Tanya Clement at the University of Texas–Austin), enable humans to create and label audio clips. Though such programs streamline

[72] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "High-level feature detection from video in TRECVID: a 5-year retrospective of achievements," in *Multimedia content analysis*, pp. 1-24. Springer, Boston, MA, 2009, 2.

the labeling process, they still require a lot of time and data; in order for the machine to identify speakers by name, humans had to label 400 to 500 one-second audio clips.[73]

Bridging the semantic gap requires significant training. A machine that returns video of a fireplace for a search for the word "fire" has not only identified the low-level features (for example, red and yellow color tones, flickering movement, and a pear-like shape), but it has assembled these features into a semantic concept narrowed down from other semantic concepts that may be similar in appearance, such as wildflowers in a windy field. In addition, some semantic concepts are more complex than others; a tennis racket is a semantic concept made up of many low-level image features that a machine must learn to recognize, but a tennis game requires identification of the tennis racket along with tennis balls, courts, and human players and audiences, among other high-level semantic concepts.

I. Audio feature analysis

Audio can be mined to identify a number of different low-level qualities, including tempo, speed, and pitch.[74] Audio's low-level features are often directly applicable to the queries humans might make of audio recordings (is someone speaking or singing? How often did this composer work in minor keys? Does this recording have applause?), making it a fruitful frontier for semantic exploration. With enough labeled data, low-level features can define and identify medium-level concepts, such as a single speaker or a specific noise.

---

[73] Tanya Clement, "Introducing the HiPSTAS Audio Toolkit Workflow: Audio Labeling," *High Performance Sound Technologies for Access and Scholarship (HiPSTAS)* (blog), University of Texas–Austin, https://blogs.ischool.utexas.edu/hipstas/2017/08/31/introducing-the-hipstas-audio-toolkit-workflow-audio-labeling/

[74] Sharon Webb, Chris Kiefer, Ben Jackson, James Baker, and Alice Eldridge, "Mining oral history collections using music information retrieval methods," *Music Reference Services Quarterly* 20, no. 3-4 (2017): 168-183, 171.

Open-source software and freeware to label and process audio has become more widely available over the last decade. One example is the machine learning tool ARLO, or Adaptive Recognition with Layered Optimization, which began as software to analyze bird calls and which has been further developed for use with humanities materials by the HiPSTAS team of researchers at the University of Texas–Austin.[75] ARLO extracts low-level audio features including "pitch, rhythm and timbre," which it then translates into spectrogram images. It then uses those images to detect clusters of patterns in an unsupervised learning process.[76] The HiPSTAS team applied this technology to identify sounds such as laughter, applause, pauses, and needle drops, as well to distinguish between dialect pronunciations.[77][78]

II. Complex semantic concepts in audio

If low-level features can be analyzed in combination to define specific sounds or voices, then these identifications can be incorporated together to identify far more complex semantic concepts. Pure audio offers less information than fully audiovisual materials, and is particularly more difficult to analyze without relying on transcripts. Analysis of audio without supplied parameters or metadata cannot always extract complex concepts such as mood, location, time period, or genre.[79]

Semantic analysis can take place within a set of low-level parameters—that is, humans can define semantic categories by their constituent sounds, and sort recordings based on these lower-level audio features. For example, a tool can be trained to distinguish between radio programs, speeches, and rallies based on the presence and quality of applause in an audio recording.[80] Structured broadcast

[75] Clement et al, "High Performance Sound Technologies for Access and Scholarship" (white paper), 5-6.
[76] Clement et al, "High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities," 3.
[77] Clement et al, "High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities," 6.
[78] Clement et al, "High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities," 6.
[79] Sander Dieleman, "Recommending music on Spotify with deep learning" (blog post), http://benanne.github.io/2014/08/05/spotify-cnns.html
[80] Clement et al, "High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities," 6.

television also offers opportunities for audio analysis; for example, loud background music usually signifies a commercial, while crowds cheering are typically a sign of sports footage and an anchor segment will be largely speech.[81]

On the other hand, humans may choose to tag categories at a high level and ask the machine itself to break down these examples into their constituent parts. For example, the Internet radio service Pandora employs humans to annotate songs to describe musical genres and qualities—high-level concepts—as well as using machine learning to make connections at lower levels that humans cannot pick out or articulate.[82]

III. Video feature analysis

Video is a multimodal medium, which means that it draws on multiple experiences (visual and aural). Video's multimodal nature makes it far more complex than audio. However, it also benefits from the ability to break down and analyze each of its modalities, in particular its visual component. Much more information can be gleaned about a video scene from its component frames than from a single frozen moment of audio. Breaking down video into still images thus offers a way for machine learning models to sidestep its time-based nature—a win for simplicity, though a reduction of the multiple dimensions of video.[83] It is far more difficult to compute features over multiple frames, and tools attempting to do so produce correspondingly less sophisticated analyses.

---

[81] Zhu Liu, Yao Wang, and Tsuhan Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI signal processing systems for signal, image and video technology* 20, no. 1-2 (1998), 64.
[82] R. Miotto and N. Orio, "Accessing Music Digital Libraries by Combining Semantic Tags and Audio Content," in *Italian Research Conference on Digital Libraries,* 26-37 (Berlin: Springer, 2011).
[83]Daniel Rothmann, "What's wrong with CNNs and spectrograms for audio processing?," *Towards Data Science* (blog), Towards Data Science Inc., 25 Mar. 2018, https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd

Video analysis starts with low-level video features including color, luminosity, edges (outlines of shapes), and texture.[84] Unlike audio, these characteristics are not usually adequate for information retrieval on their own. (Some exceptions include corpuses of video meant for advertising, art, or film analysis, in which cases color may be a defining attribute.) However, they may be used to detect shot boundaries, or cuts between a video's component sequences. Most shot boundary detection methods identify sequences by identifying abrupt cuts, such as sharp differences in color histograms from one frame to the next.[85] The TRECVID ("Text REtrieval Conference: Video") series of forums, sponsored by the National Institute of Standards and Technology, has sponsored experiments in video feature analysis since 2001, focusing on different semantic concepts and corpuses of data each year. In these worldwide forums, research groups developed competing models to identify and label sequences in videos whose genre ranged from news broadcast to ephemeral and orphan media supplied by the Prelinger Archives.[86] Breaking down videos shot by shot allows for more precise tagging and description.

Analysis of low-level features in combination can also yield semantically complex features like faces and objects. Facial recognition is a machine learning tool that succeeds based on two facts: every human's face is unique and therefore viable as an identifying tool, and the presence of a human face (a generic and predictable set of features) is easy for machines to recognize.[87] Facial recognition is often used to identify specific individuals, but it can also be used to identify emotions, age, and expressions.[88] Facial identification of individuals raises serious privacy issues, as their success relies on access to images

[84] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert, "Feature Visualization: How neural networks build up their understanding of images," *Distill* 2, no. 11 (2017): e7, https://distill.pub/2017/feature-visualization/
[85] Smeaton et al, "Video shot boundary detection: Seven years of TRECVID activity," 414.
[86] Alan F. Smeaton, Paul Over, and Aiden R. Doherty, "Video shot boundary detection: Seven years of TRECVID activity," *Computer Vision and Image Understanding* 114, no. 4 (2010), 411-412.
[87] Jay D. Aronson, "Computer Vision and Machine Learning for Human Rights Video Analysis: Case Studies, Possibilities, Concerns, and Limitations," *Law & Social Inquiry* 43, no. 4 (2018), 1200.
[88] "Amazon Rekognition," *Amazon Web Services,* Amazon Web Services, Inc., 2019, https://aws.amazon.com/rekognition/

of the person, and private citizens are unlikely to have given consent for this use of their likeness. For this reason, facial recognition tools are most suitable for celebrities and others in the public eye.[89] In addition, such tools suffer from limited datasets; many have been trained on data skewed heavily toward white faces, and are correspondingly bad at recognizing black faces (both as individuals and as human faces at all). Heavy facial hair and head coverings may make identification impossible no matter who the person is.[90]

Object recognition relies on classification techniques similar to those used for faces, but the range of successful answers is less specific and more generic. A class of object such as a tennis racket, a cat, or an airplane is expected to be identifiable across collections in essentially unbounded ways; a machine learning model must be trained to recognize an airplane whether it's small or large, blue or red, in the sky or on the ground, or an old-timey flying machine or a modern jet. It is also significantly more difficult to identify the exact outlines of the airplane within the image than it is to simply say whether or not there is an airplane somewhere in the image, and the former is much more important for tracking objects across multiple frames of video.[91] Despite these challenges, large-scale projects and competitions such as TRECVID and the ImageNet Large Scale Visual Recognition Challenge[92] have concentrated efforts in this area.

IV. Complex semantic concepts in video

Objects, faces, and audio can be combined to yield complex semantic concepts such as "events," which are compound activities ranging from explosions to birthday parties. Like object

[89] "Amazon Rekognition," *Amazon Web Services,* Amazon Web Services, Inc., 2019, https://aws.amazon.com/rekognition/
[90] Aronson, 1201-1202.
[91] Russakovsky et al, 2.
[92] "The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)," *ImageNet,* Stanford Vision Lab, 2017, http://www.image-net.org/challenges/LSVRC/

recognition, event recognition is complicated by the range of possible representations of a single semantic concept. In identifying a tennis game, for example, a machine learning model can encounter a large potential set of false positives (squash also involves rackets and balls; tennis commercials might incorporate scenes on a court).[93]

In highly structured video, events can also be identified in simpler ways. For example, shot boundary detection is particularly useful in news broadcasts, which can be separated into predictable shot categories such as news anchors, field segments, sports coverage, and weather reports. Such a breakdown lends coherent meaning to each story unit, and can be used to index and access material.[94]

V. Hashing images

Though much audiovisual processing for descriptive content is complex, they almost exclusively operate by breaking moving images down into their component still images. Still images taken in isolation represent a video's visual content but are not tied to a specific length, edit, or version of a video. This versatility can be leveraged to find matching content by hashing.

There are two major types of hashing: cryptographic and perceptual. Cryptographic hashing is familiar to archivists from the concept of digital fixity. In cryptographic hashing, an archivist runs an image through a standard algorithm to produces an output called a checksum, which is a string of letters and numbers. If the archivist runs an identical image through the same algorithm, they will receive the same checksum; if a single byte is changed in the image, the resulting checksum will be very different. Cryptographic hashing is a useful tool to monitor whether an image has changed during its

[93] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al, "Imagenet large scale visual recognition challenge," *International journal of computer vision* 115, no. 3 (2015), pg. 6

[94] Tat-Seng Chua, Shih-Fu Chang, Lekha Chaisorn, and Winston Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *Proceedings of the 12th annual ACM international conference on Multimedia*, ACM, 2004, 656.

time in digital storage, as checksums that do not match are clear signs that something has happened to the file. Checksums can also identify duplicates across collections, and may be used as a flag for weeding or copy cataloging (depending on the policies of the institution).

Perceptual hashing is similar to cryptographic hashing, but the algorithms used to generate the output (in this case, a "fingerprint") standardize the input file before analysis. For example, all audio files might be transformed to a standard sample rate or codec. This change allows the perceptual hash to identify materials that are almost the same, but have slight differences, such as an access derivative of a higher-quality master or the same video with and without bars and tone at the beginning. Perceptual hashes can therefore be put to much more versatile search uses than cryptographic hashes, as perceptual hashing can reveal similar sequences across very different videos. For example, perceptual hashing might reveal that the same news clip was used on the nightly news across multiple broadcast networks, while cryptographic hashing would only be able to tell that each nightly news program was different as a whole.[95]

C. Turning audio and video into text

As audiovisual archivists well know, audiovisual materials are extremely difficult to search for and discover within a typical collections management system. The time-based nature of audio and video requires significant time investment on the part of both the archivist and the researcher. There are few ways to automatically tag audio and video within catalogs, and even fewer ways to "skim" audiovisual materials if they miraculously pop up in a researcher's search results. And unlike a pixel of video (out of which objects and faces are built) or a moment of audio (out of which sounds are built),

[95] Andrew Weaver, "Adventures in Perceptual Hashing" (blog post), AAPB NDSR blog, National Digital Stewardship Residency, 20 Apr. 2017, https://ndsr.americanarchive.org/2017/04/20/adventures-in-perceptual-hashing/

words—the analogous breakdown of text—have intrinsic meaning in or out of context, and yield far more meaningful raw information.[96] Automated and AI-based tools are generally far more advanced for text-based analysis, because digital humanities research has focused on tools to parse, summarize, and detect patterns in textual output.[97] This body of work may provide directions for analysis of transcripts and other large blocks of text. Machine learning offers methods to transform audiovisual materials into text and thereby simplify discovery within today's catalogs—and even put audiovisual content on the same footing as textual and manuscript documents.

I. Speech-to-text transcription

Speech-to-text algorithms render spoken language as text, generating transcriptions of audio without accompanying written records. Speech-to-text, also known as automatic speech recognition (ASR), is a fundamental part of commercial speech assistants such as Siri, Alexa, and Cortana, which translate spoken commands into words, sentences, and meaning.[98] Speech recognition differs from audio analysis in that it focuses on linguistic content instead of sounds and features. It overlaps with, but is not the same as, technology that recognizes individual speakers' voices; speech-to-text tools can be speaker-dependent, meaning that they are only consistently accurate when interpreting a single voice, or speaker-independent, meaning that they can handle a wide range of accents and syntaxes.[99]

In order to translate audio to text, words must be recognized in context. Early speech-to-text tools tried to segment speech into phonemes and identify each word individually, a process that was easily disrupted by homonyms such as "threw" and "through." Machine learning methods called

---

[96] Lauren Tilton and Taylor Arnold, "NEH Grant Narrative: Distant Viewing," *Distant Viewing,* University of Richmond/National Endowment for the Humanities, 2018, https://distantviewing.org/pdf/neh_grant_narrative.pdf

[97] Winfred Phillips, "Introduction to Natural Language Processing," *Consortium on Cognitive Science Instruction,* Illinois State University, 2006, http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php

[98] Xuedong Huang, James Baker, and Raj Reddy, "A historical perspective of speech recognition," *Communications of the ACM* 57, no. 1 (2014), 103.

[99] Huang et al. 99.

hidden Markov models have enabled speech-to-text tools to take into account the statistical likelihood of a word occurring at that point in the sentence, allowing semantic analysis to occur simultaneously with phonetic analysis.[100] Machine learning also allows tools to train with recordings from speakers with diverse accents and varying recording quality.[101]

These advances in speech-to-text methods have propagated a raft of commercially-available apps, many more so than the other machine learning applications discussed in this section. The demand for speech-to-text stems from the massive medical and legal fields, where doctors and lawyers rely on medical transcriptionists and court reporters, as well as from interviewers in journalism and academia.[102] Most of these tools are commercial, or have been acquired by commercial companies (for example, Pop Up Archive, an NEH-funded project acquired by Apple and subsequently taken offline).[103] Notable tools that have been made available as open source projects include Kaldi (begun at a Johns Hopkins University workshop)[104] and CMUSphinx (Carnegie Mellon University),[105] among others.

While transcripts can be generated for videos using only the audio track, some speech recognition methods incorporate visuals to aid comprehension. Systems that build in lipreading ability have been demonstrated to improve intelligibility in loud rooms or in low-quality audio. However, these systems can perform worse than audio-only speech recognition in clean audio environments, as it

---

[100] Huang et al, 97.
[101] Huang et al, 99.
[102] "13 Startups Transcribing Voice to Text Using AI," *Nanalyze.com*, Nanalyze, 29 July 2018, https://www.nanalyze.com/2018/07/voice-to-text-transcribing-ai/
[103] Jennifer Howard, "Pop Up Archive Filled a Need for Audio Archiving, and Apple Noticed," *Humanities: The Magazine of the National Endowment for the Humanities* 38, no. 4 (Fall 2017), National Endowment for the Humanities, https://www.neh.gov/humanities/2017/fall/feature/pop-archive-filled-need-audio-archiving-and-apple-noticed
[104] "History of the Kaldi project," Kaldi documentation, Kaldi, accessed 2 May 2019, http://kaldi-asr.org/doc/history.html
[105] "About CMUSphinx," CMUSphinx Open Source Speech Recognition Toolkit, CMUSphinx, https://cmusphinx.github.io/wiki/about/

is difficult to train a multimodal (audio and video) framework when to favor one feature over another.[106]

II. OCR of static graphic text

Text can also be extracted from video through optical character recognition. Tasks involved in extracting this text include text detection, or determining whether there is text in the scene; text localization, or determining where the text is in the scene; and text information extraction, or the process of making it readable for the machine.[107]

Static graphic text, such as chyrons, on-screen captions, and credits, is straightforward to extract once it is identified as present within the image. This type of video text is often consistent—for example, credits are typically white text on black backgrounds, and broadcast graphics with text are found in the lower third of the screen. Text that appears in the exact same location within the image every time (and within images of the same resolution every time) can be extracted with extremely simple scripts; for example, Waldo Jaquith of Richmond Sunlight uses a series of open source tools to crop bill and legislator names from screenshots of Virginia General Assembly footage, convert the cropped names to black and white, OCR and spell-check them.[108] But in many other materials, the position of graphic text will vary slightly from program to program, file to file, or even master to derivative (the position of a chyron, as measured in pixels, will not be the same in a 4K original file and an HD streaming file). Scrolling credits and chyrons change position from frame to frame, which requires systems to decipher partial characters and piece together sentences. Some algorithms that

[106] Fei Tao and Carlos Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26, no. 7 (2018), 1286.
[107] Qixiang Ye and David Doermann, "Text detection and recognition in imagery: A survey," *IEEE transactions on pattern analysis and machine intelligence* 37, no. 7 (2015), 1480.
[108] Waldo Jaquith, "How I OCR hundreds of hours of video" (blog post), 10 Feb. 2011, https://waldo.jaquith.org/blog/2011/02/ocr-video/

translate images to greyscale for processing can even be foiled by non-contrasting background and text colors.[109]

III. OCR of variable scene text

OCR of variable scene text, such as signs, license plates, and billboards, is far less distinguishable from its surrounding scenery than more static graphic text. Unpredictable colors and lighting, as well as moving images behind text, reduce the ability of machines to recognize and translate text.[110] Other variables include the width of the text box (ranging anywhere from a license plate in the distance to a close-up of journal entry), skewed and angled text, and inconsistent fonts.[111]

In both static and scene text extraction, the most efficient way to address these challenges is correctly detecting the areas of the video frame that actually hold text, as this cuts down on the noise when OCR is used for non-textual information. Commercial OCR tools are advanced enough to be widely reliable when fed with readable information, with success rates of more than 99 percent when fed with scanned text.[112] Researchers from the University of Mannheim, who have developed a method of identifying and isolating textual regions in video, can give these image files to "any standard OCR software" for results.[113]

---

[109] Rainer W. Lienhart and Frank Stuber, "Automatic text recognition in digital videos," in *Image and Video Processing* IV, vol. 2666, International Society for Optics and Photonics, 1996, 186.
[110] Lienhart and Stuber, 181-182.
[111] Ye and Doermann, 1482.
[112] Ye and Doermann, 1480.
[113] Lienhart and Stuber, 187-188.

2. Tools for access

Though research into image, audio, and video analysis has grown as machine learning models have become more sophisticated, it remains unusual to access such results through a non-textual database. Some scholars have begun to explore "sensory" databases, in which imagery, sound, and movement provide a nexus of exploration.[114] Exploring by similarities is a relatively new method of discovery, but it is also a method of analysis that surfaces patterns; for example, the "Neural Neighbors" project at Yale's Digital Humanities Laboratory provided evidence of the tropes and technical limitations of 19th century portraiture. Such methods usually rely on machine learning to identify the components of an image and calculate the similarities in images across collections.[115]

Natural language processing may also provide more generalized and flexible methods of access for textual content. Keyword searches currently depend upon locating instances of an exact phrase, but natural language processing offers opportunities for more contextual and unambiguous searches. [116] For example, semantic analysis can map relationships between between sets of concepts, can expand a net of related concepts (for example, expanding a search for "jaguars" to other big cats) at the same time as it weeds the field of false positives (for example, knowing to exclude Jaguar cars from a search for "jaguars").[117] Digital humanities tools that parse, summarize, and detect patterns in textual output may be used to offer more points of discovery for transcripts.[118]

Machine learning may also be leveraged in the moment of research, relieving the archivist of anticipating all their users' needs and artificially limiting research inquiries based on pre-processed

---

[114] Barbara Flueckiger, "A Digital Humanities Approach to Film Colors," *Moving Image: The Journal of the Association of Moving Image Archivists* 17, no. 2 (2017), 71-94.
[115] "Neural Neighbors: Capturing Image Similarity," *Digital Humanities Laboratory*, Yale University Library, accessed 3 May 2019, http://dhlab.yale.edu/projects/neural_neighbors.html
[116] Phillips.
[117] Pradeep Palani, "Understanding Semantic Analysis (And Why This Title is Totally Meta)," *Zeta Global,* Zeta Global blog, 10 Jan. 2018, https://zetaglobal.com/blog-posts/understanding-semantic-analysis-title-totally-meta/
[118] Phillips.

materials. For example, if a researcher who approaches an archivist with a defined research interest, such as a certain person or audio quality, the archivist may be able to train new classifiers that target those research interests at the moment of inquiry. Such a method would be possible in relatively short amounts of time if the audiovisual corpus is "pre-processed," or already subjected to (in this example) basic facial detection and audio waveform analysis; once a specific face or sound has been identified, it is simple to find it throughout the digital materials.[119]

If audiovisual content is to be described in such detail, effective access will depend on addressing familiar limitations of content management systems. For instance, metadata and tags for audiovisual content often must be associated with an entire audiovisual file rather than a time-stamped moment. In such systems, if a tool recognizes Barack Obama's face in a news broadcast, the video file can be tagged with "Barack Obama." However, if Barack Obama only appears at the 5:00-minute mark in a 20-minute video, it would be more meaningful to retain the relationship between the image of Obama and its appearance in the audio or video file, eliminating the need to scrub through or watch the entire file to find his appearance. Time-based tagging can be achieved at a low level by recording the timestamp of Barack Obama's appearance in a container like a text file, or at a higher level by dynamically tagging Barack Obama's face in the footage. An application geared towards audio and video, such as the Oral History Metadata Synchronizer, developed by the Louie B. Nunn Center for Oral History at the University of Kentucky Libraries, can allow transcripts to be synchronized with the audiovisual content, whether word by word or in larger indexed sequences.[120]

---

[119] Parkhi et al, cited in Tinne Tuytelaars, "Content-based analysis for accessing audiovisual archives: alternatives for concept-based indexing and search," in *2012 13th International Workshop on Image Analysis for Multimedia Interactive Services,* IEEE, 2012, 2.

[120] "Oral History Metadata Synchronizer," Oral History Metadata Synchronizer, *Louie B. Nunn Center for Oral History at the University of Kentucky Libraries*, 2019, http://www.oralhistoryonline.org/

III. Integrating machine learning into archival workflows

While automation and machine learning can address many of the same problems in archival description, the former is a much simpler concept than the latter. Automation is powerful, and its exact workings can be difficult for beginners to grasp. But despite these qualities, it is also ultimately transparent, self-descriptive, and directly modifiable by humans who must take responsibility for the final output. The mechanics and nature of machine learning, a complex and nascent technology with low literacy and regulation, require consideration on their own.

Though this paper addresses applications of automated tools, a review of the manners in which they may be integrated into archival workflows is ground that has been covered by many in the context of digital asset management systems, as well as microservice-based workflows. The following section will primarily address the incorporation of artificial intelligence and machine learning into local workflows.

1. Software, tools, and infrastructure

What does it look like to incorporate machine learning into archival workflows? The answer varies between tool, repository, and existing infrastructure, but in each case requires largely the same considerations as the implementation of any other processing tool or workflow: selection of a tool that fits into the local digital processing ecosystem, and close collaboration with technical support staff.

It is worth noting that most large technology companies have separate research (noncommercial) and commercial divisions, but this division is muddy, as the parent company owns all the intellectual property developed by the research unit.[121]

A. Commercial products

Incorporating machine learning into archival workflows does not have to involve any technical literacy. In fact, many commercial machine learning tools are presented as complete packages and aimed at beginners. They are available through standalone applications or APIs, and can be used either as a service (bought as a subscription or on a per-use basis, and as the responsibility of a cloud computing service) or as a product (bought in full and installed by the user).

Many cloud computing companies offer machine learning "as a service," in a model that provides customers with access to software and tools that are hosted and run by the companies on their backend. Cloud computing—that is, a framework that allows remote and on-demand access to community computing resources—allows the customer to configure and use the offered services without the technical overhead of installing and maintaining them. Machine learning as a service, or MLaaS, is an appealing option to institutions that want to try out machine learning tools without having to power them locally. The tradeoff of this model is that the customer also gives up ultimate control of the computing resources.[122]

MLaaS providers supply a range of machine learning tools to train and evaluate models, often using built-in algorithms. The customer provides the data and, depending on the provider, anywhere from a novice to an expert's understanding of machine learning, as the amount of automation (as

[121] Solon.
[122] Peter Mell and Tim Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology, Special Publication 800-145 (2011).

opposed to manual tinkering) varies wildly.[123] It is often effectively marketed to institutions whose materials are already held by a cloud storage provider; Amazon Web Services, Google Cloud Platform, and Microsoft Azure, among others, offer machine learning applications by subscription or pay-as-you-go, by the hour or by the result.[124] The significant downside to this model is that the tools, whether in the form of an API or a graphical user interface, only work with materials uploaded to storage owned by the MLaaS company. For the tools to work with collections on local storage, they must also be hosted locally—which would make them no longer an on-demand "service."

Machine learning applications bought and hosted in-house are aimed at teams of developers and data scientists, as they require far more expertise to mount and run, as well as much larger hardware investments for adequate speed and power. Because of this, they are not able to scale up according to demand in the same manner as the MLaaS model; however, they offer far better control over the training and results.[125]

Whatever service is used, care should be taken when integrating machine-generated metadata into a digital asset management system. Tags, descriptions, and other metadata created by machine learning services should be kept separate from human-generated metadata and tracked as a user to make it easy to isolate the service's data, audit its quality, and toggle it on and off in search results. In addition, tags should be reported with confidence levels, or the percentage of certainty the machine had when assigning the tag, and users should have the ability to filter at whatever confidence level they desire.[126]

---

[123] Janakiram MSV, "An Executive's Guide To Understanding Cloud-based Machine Learning Services," *Forbes.com,* 1 Jan. 2019, https://www.forbes.com/sites/janakirammsv/2019/01/01/an-executives-guide-to-understanding-cloud-based-machine-learning-services/#52d5709e3e3e
[124] Janakiram MSV, "An Executive's Guide To Understanding Cloud-based Machine Learning Services," *Forbes.com,* 1 Jan. 2019, https://www.forbes.com/sites/janakirammsv/2019/01/01/an-executives-guide-to-understanding-cloud-based-machine-learning-services/#52d5709e3e3e
[125] Rob Light, "Machine Learning as a Service (MLaaS)," *G2 Digital Trends* (blog), G2 Crowd, accessed 5 May 2019, https://blog.g2crowd.com/blog/trends/artificial-intelligence/2018-ai/machine-learning-service-mlaas/
[126] "6 Best Practices for Implementing AI in a DAM," *MediaValet* (blog), Mediavalet, accessed 5 May 2019, https://www.mediavalet.com/blog/implementing-artificial-intelligence-in-dam/

B. Open source tools

Free and open machine learning tools are available in a range of models, from snippets of code on GitHub pages to well-documented, grant-funded toolkits created within libraries and archives. Some open source projects have attracted significant community attention and active development and maintenance,[127] while others exist as fully developed end products, as may be the case with grant-funded projects with a firm end date and deliverables.[128] Many more projects exist as simple snippets of code that live on GitHub without documentation, context, or community backing; the inverse of this state are academic papers describing successful projects in detail, with no or little corresponding code. A significant amount of time must therefore be reserved to testing out code and piecing it together, especially in relatively newer fields of AI such as computer vision (as opposed to natural language processing, with much better availability of libraries and working code).[129]

Open source projects often require high start-up costs from an institution, or at least a baseline level of technical comfort with setting up and troubleshooting unknown software. Poorly documented code in particular requires a lot from its users, though even well-documented code may be difficult to understand or implement for machine learning novices. In addition, open software rarely exists as an API or plug-in, given the strain on resources and development time that represents for a small-scale project. This means that the metadata they generate will be likely kept separate unless their output is automated to pass to local databases or storage—a helpful way to track and separate AI-generated metadata, but another point within the process that requires intervention for full

[127] See, for example, the popular GitHub repository for "OpenPose," a computer vision project under the aegis of the Perceptual Computing Lab at Carnegie Mellon University: https://github.com/CMU-Perceptual-Computing-Lab/openpose
[128] See, for example, the suite of tools offered through the NEH- and IMLS-funded HiPSTAS project at the University of Texas–Austin: https://github.com/hipstas
[129] Conversation with Lauren Tilton and Taylor Arnold.

integration. Despite these potential roadblocks, there remains one major benefit to implementing an open source tool: the ability to modify and build upon the existing code, for those willing to invest the time to become comfortable with the process.

C. Developing tools

Machine learning tools may be trained on local datasets for use with more specific types of collections. For example, a regional archive seeking to generate transcripts for its oral histories might hope to train its tools on the local accent, for more accurate speech-to-text. Typically, a user will choose to train on top of a pre-existing generic model, rather than attempt to build one from scratch.

Many commercial products tout the ability to customize their machine learning services by uploading labeled datasets to the company's servers and clicking "generate" to train a custom model. For example, Microsoft Azure's Speech Services allow the user to use training to address misrecognized proper nouns or accents,[130] while Amazon Transcribe advertises the ability to add new words to a custom model.[131] There are significant limitations to these features, most importantly the fact that the user has little control over the choice of model, parameters, or the ability to run these models outside the cloud environment.[132] In addition, there are usually separate pricing models for customization that often operate per-use, so the user must be prepared to incur extra costs beyond the cost of the tool itself.

Open source tools also offer the ability to customize models, with the added benefit of being able to run these models locally. However, as with initial installation, extending open source models

[130] "Train a Model for Custom Speech," Microsoft Azure documentation, *Microsoft,* 1 May 2019, https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-train-model
[131] "Amazon Transcribe," Amazon Web Services, Amazon, accessed 5 May 2019, https://aws.amazon.com/transcribe/?n=sn&p=r
[132] Grigorev.

usually relies on a comfort level (and investment of time and hardware) that goes beyond the graphical user interface-based workflows of a commercial product. And while the open source model is generally attractive for its adaptability and customization, open source machine learning has a similar limitation to commercial services—the user has little choice of model or parameters, unless they are willing to train their own model from a more basic starting point. Unlike most commercial products, however, open source tools do reward users who have (or acquire) enough knowledge to exercise control over these options.

It is worth noting that even tools developed apart from a commercial product base are almost never built "from scratch" or based on completely custom datasets. Machine learning development relies heavily on algorithms already partially trained on existing generic datasets, because the work it would take to train entirely new models is prohibitively expensive—and limiting. If an object recognition model cannot "see" the landmarks around Brooklyn in a home movie collection, there are two options. The first is to train the existing model to see the Coney Island Parachute Jump and the Williamsburgh Savings Bank as well as generic cars and apartment buildings, for example by ignoring the last few layers of the model's neural network and building a simpler, Brooklyn-based model on top of the rest. The other option is to train a model with a Brooklyn-only dataset. However, because it lacks generic images, the model is likely to return false positives, for example seeing the Williamsburgh Savings Bank in every clock or tall building. The way to fix this problem would be to incorporate a generic dataset—at which point one should just save the time and work with an existing tool already trained on this dataset. It is the accumulation of knowledge that creates a successful model, not just the knowledge geared towards the local collection.[133]

[133] Conversation with Lauren Tilton and Taylor Arnold.

Developing or refining open tools for local use demands collaboration between practitioners who understand its technical implementation, and those who understand its place in the description workflow. One example of a successful partnership is that of the Distant Viewing Lab, directed by Lauren Tilton, Assistant Professor of Digital Humanities, and Taylor Arnold, Assistant Professor of Statistics, both at the University of Richmond. The lab, funded by an NEH grant, develops tools to extract metadata features from video, and uses them to conduct computational research on moving image culture. The toolkit is meant for reuse and is available on GitHub with tutorials and documentation. The tool's genesis and theoretical underpinnings draw heavily on digital humanities context and knowledge, but its execution relies on statistical and technical knowledge. Tilton and Arnold describe their work as a rapid-fire partnership, noting that it would be unrealistic to expect either to master every dimension of the project, but that neither facet succeeds without mutual affinity and constant back-and-forth refining and critiquing.[134]

D. Integrating tools into workflows

Before investing in machine learning tools, it is useful for an archives to understand its own goals. The choice of machine learning tools, and their usefulness, depends on the needs of the staff and users. If archivists want to make a hidden collection of oral histories more accessible, then using automatic speech recognition to generate transcripts is a bounded and achievable goal. If users don't want to ask computational questions, then maybe tasks like scene segmentation are not worth the investment.

[134] Conversation with Lauren Tilton and Taylor Arnold.

Because machine learning tools can only be trained to perform one task, they are generally integrated into workflows as "microservices," or small tools that perform a single, discrete task. These microservices may be applications with graphical user interfaces operated manually by humans, such as the browser-based interaction offered by Google Cloud AutoML Vision. They may be APIs or command line tools that can be coded into automated workflows on the client's side.[135] Or they may be APIs that are integrated into large digital asset management systems, an option that requires the support of the company or individuals who maintain the DAMS.[136]

Microservices gain utility when they are strung together in a pipeline and automated. (In this way, machine learning services and automation experience crossover.) For example, a team at Brandeis and Vassar have developed a toolkit for the WGBH Library and Archives that will allow archivists to drag-and-drop media files into a system of speech-to-text, entity recognition and extraction, metadata normalizing, and automated ingest.[137] Such setups are impossible to automate if work only takes place in a graphical user interface, though a monolith environment like a DAMS may be able to incorporate similar features.

E. Using collections as datasets

When used to train a machine learning model, an archives' collections begin to serve as datasets. Often this way of thinking raises a number of logistical questions that envision digital materials in a new light.

---

[135] Google Cloud, "Making an online prediction," *AI & Machine Learning Products documentation,* Google Cloud, 17 Apr. 2019, https://cloud.google.com/vision/automl/docs/predict

[136] Martin Wilson, "AI in DAM: The Challenges and Opportunities," *Digital Asset Management News*, accessed 5 May 2019, https://digitalassetmanagementnews.org/features/ai-in-dam-the-challenges-and-opportunities/

[137] James Pustejovsky, Nancy Ide, Marc Verhagen, and Keith Suderman, "Enhancing Access to Media Collections and Archives Using Computational Linguistic Tools," in *CDH@TLT*, 2017, 24-25.

For example, collections are often inventoried for storage space or numbers of items; in machine learning, these amounts of data become important to gauge whether a successful machine learning implementation is possible. How much data is required varies depending on the level of training being done; to train a computer vision deep learning application from scratch, for example, requires millions or even billions of images, far beyond the number of images in a typical digital asset management system.[138] However, training a model to recognize a new voice or object—on top of an existing tool—requires far less. The HiPSTAS team at University of Texas–Austin found that training a tool to recognize Terry Gross, host of NPR's *Fresh Air,* required labeling approximately 400-500 clips of her voice.[139] Google research scientist Ashok Papat has suggested that OCR models need 1,000 lines of transcribed text to recognize new typefaces or layout.[140] In order to avoid "overfit," or over-sensitivity to the particular quirks of the training dataset, archivists must hold back some of the training dataset to use instead as testing data. It is impossible to test the success of an algorithm on the data it has already seen, even in a very large corpus, given the fact that machine learning's advantage should be its ability to classify complex and surprising examples.[141] However, holding back test data should not come at the price of making the training dataset too small; overfitting can also occur when the model is not fed enough data, as the variance in a small dataset can often be essentially equivalent to an algorithm that relies on total randomness.[142]

Creating even these smaller datasets requires extensive human time, even in an institution with good metadata practices. Many archival institutions have adopted structured metadata practices and seem like a natural source of data and metadata for machine learning; for example, an audio archive of

---

[138] Wilson.
[139] Tanya Clement, "Introducing the HiPSTAS Audio Toolkit Workflow: Audio Labeling."
[140] Cited in Bethany Nowviskie, "Reconstitute the World," Bethany Nowviskie (personal website), 12 June 2018, http://nowviskie.org/2018/reconstitute-the-world/
[141] Domingos, 79.
[142] Domingos, 80.

lectures tagged with the names of the speakers seems like a fruitful place to start in performing voice recognition. But even seemingly simple lectures require more processing work. Applause, interstitials, and laughter must be excised; multiple speakers must be identified and then separated into their constituent voices; and audio samples must be standardized into identical formats and codecs before processing can happen. Inconsistent metadata practices may also make the data unusable in machine learning contexts (as in the situation of the Teenie Harris project at the Carnegie Museum of Art, where photographs had been titled differently according to decades of cataloging standards), and must be standardized or finalized. In addition, institutions have rarely worked through their backlog of materials. It is unusual for entire collections to be already tagged with standardized metadata, which means that institutions are either limited to already-prioritized collections or responsible for doing the work of cataloging prior to embarking on machine learning workflows. A subcategory of machine learning services and open source tools deals specifically with ways to automate the process—whether through the human labor that Google offers to label users' data[143] or in auxiliary tools such as the HiPSTAS project's Audio Labeling Toolkit, used to facilitate the actual labeling of Terry Gross sound clips.[144] Dataset creation is rarely considered a distinct scholarly category in itself, although it is essential to the success of the training process and representative of serious intellectual effort.[145]

Another logistical question concerns the quality of the data. Low-quality audio and video information undermines the ability of a model to work, either by obscuring information from the model when it is run, or by training the model on noisy, low-quality data and interfering with its ability to recognize information at all. Some common sources of noise in data come from analog audio

[143] "Human labeling," *AI & Machine Learning Products documentation,* Google Cloud, 15 Nov. 2018, https://cloud.google.com/vision/automl/docs/human-labeling
[144] Tanya Clement, "Introducing the HiPSTAS Audio Toolkit Workflow: Audio Labeling."
[145] Jessica Otis, @jotis13, Twitter post, 9 Mar. 2019, https://twitter.com/jotis13/status/1104410030458265600

and video artifacts, such as skewing or tearing in analog video or the sounds of pops, hisses, and scratches in audio playback.[146] Other sources of noise are introduced during digitization (for analog media) or the process of creation (for born-digital media), such as digital clipping at the high and low ends of luminance levels and audio frequencies. The overall level of digital quality may also limit analysis; for example, accurate face detection requires that faces be approximately 100 pixels by 100 pixels, which means that standard definition video may be adequate to identify faces in close-up, but that long shots will not be identifiable unless the video is ultra high definition.[147] In other cases, shaky handheld footage or ambient noise will limit a user's ability to analyze audiovisual materials.[148] Archivists and format experts are able to address some of these issues, but others are inherent in the digital files and cannot be improved without changing the qualities of the original file.

2. Advantages of machine learning

Machine learning is an extremely new technology, and one that has not yet lived up to its full potential. However, machine learning can accomplish many descriptive processing tasks that are completely outside the scope of automation. Machine learning is also a process that improves rapidly with training, with less human effort than it takes to improve automated processes. Processing large amounts of material can constitute a symbiotic relationship between the tools of machine learning and the collections themselves. Retraining and refining models according to local collections trains the models at the same time as they grant access to collections, meaning that machine learning models will continue to improve as they are used.

---

[146] Lienhart and Stuber, 186.
[147] Taylor Arnold.
[148] Jay D. Aronson, Shicheng Xu, and Alex Hauptmann, "Video analytics for conflict monitoring and human rights documentation" (technical report), Center for Human Rights Science Technical Report, Carnegie Mellon University, 2015, 4.

What follows is not an exhaustive comparison, but an attempt to enumerate some of the applications of machine learning most useful to archivists that cannot be addressed by automation.

A. Accessibility and access to audiovisual materials

Some of the most advanced applications of machine learning are in areas that directly address core archival mandates for accessibility and access. For example, the World Wide Web Consortium's Web Content Accessibility Guidelines (version 2.0), which require captions for all prerecorded and live audiovisual content, has become a benchmark for measuring compliance with the ADA and has been adopted in phases by the U.S. federal government, whose Section 508 guidelines apply to every public institution in the country.[149] While speech-to-text is not yet adequate for professional broadcasting use, it can provide a solid starting point for archives making their audiovisual content available online. With the industry standard for transcription set at about four hours of transcription time for one hour of speech—and more for low-quality audio—a solution for transcription is crucial for archives that cannot afford to hire full-time transcriptionists.[150] Speech-to-text applications are estimated to transcribe clean audio with an average eight percent error rate, which requires cleanup from archivists, but at far less than the 4:1 ratio.[151] They are also cost-effective for cash-strapped archives; an application called Trint currently charges $15 per hour of audio.[152]

Though less ready for prime time, object, facial, and event recognition can be put to use for accessibility purposes as well. Facebook, among others, uses deep learning to provide alternative text

---

[149]"Guideline 1.2 Time-based media," *Web Content Accessibility Guidelines (WCAG) 2.0*, World Wide Web Consortium, eds. Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, et al., 11 December 2008, https://www.w3.org/TR/WCAG20/

[150] "Transcription time per audio hour: How long does transcribing really take?," *Opal Transcription Services,* Opal Transcription Services, accessed 5 May 2019, https://www.opaltranscriptionservices.com/transcription-time-per-audio-hour/

[151] Jesse Jarnow, "Transcribing Audio Sucks—So Make The Machines Do It," *Wired,* 26 Apr. 2017, https://www.wired.com/2017/04/trint-multi-voice-transcription/

[152] "Pricing," *Trint.com,* Trint, accessed 5 May 2019 https://trint.com/pricing/

for images on its platform, automatically providing information to users with screen readers about what the images may depict.[153] A text description of video would be a starting place for providing accessibility (though these descriptive captions would not by themselves fulfill accessibility requirements as set out by the federal government, which requires synchronous audio descriptions of digital video,[154] and would only fulfill Level A accessibility standards as set by the World Wide Web Consortium).[155]

Fulfilling accessibility requirements has the added benefit of increasing access for all users. Because search and content management systems operate on the basis of text, having textual transcriptions, captions, and tags will surface materials that were previously impossible to find outside of the context of their collections. An archive that takes the extra step of time-stamping or synchronizing transcripts to audio and video is likely to find that users engage more directly with the audiovisual nature of the materials.[156] In a long history of archival and historical documents, text dominates research and discussion. Somewhat paradoxically, machine learning offers an opportunity to undermine the supremacy of textual documents through translating audiovisual documents into text and putting the formats on the same level.[157]

B. Describing large amounts of material

Machine learning techniques are flexible enough to be appropriate for fully processed, partially processed, and unprocessed materials. The emphasis on metadata in archives and libraries presents

---

[153] Cade Metz, "Facebook's AI Is Now Automatically Writing Photo Captions," *Wired,* 5 Apr. 2016, https://www.wired.com/2016/04/facebook-using-ai-write-photo-captions-blind-users/
[154] "Create Accessible Video, Audio and Social Media," *Section508.gov,* U.S. General Services Administration, May 2018, https://www.section508.gov/create/video-social
[155] "How to Meet WCAG 2 (Quick Reference): Guideline 1.1—Text Alternatives," *Web Accessibility Initiative,* World Wide Web Consortium, 29 Jan. 2019, https://www.w3.org/WAI/WCAG21/quickref/?versions=2.0&showtechniques=111%2C123%2C125%2C129#text-alternatives
[156] Doug Boyd, "OHMS: Enhancing access to oral history for free," *The Oral History Review* 40, no. 1 (2013), 96-97.
[157] Patricia Cohen, "Digital Keys for Unlocking the Humanities' Riches," *The New York Times,* 16 Nov. 2010, https://www.nytimes.com/2010/11/17/arts/17digital.html?_r=0

datasets that are generally useful for supervised learning, as collections that have been tagged may be ready for training (after some cleanup). For untagged and unprocessed collections, unsupervised learning, or clustering, can provide a point of entry for collections. While such surface-level descriptions and hints are not appropriate for full access, they act as a lens onto a collections that might otherwise remain opaque. Clustering relieves the cataloger of the work of an initial sort, and gives them a direction for further description.

Automation cannot provide descriptive help to archivists, because the semantic gap, while a roadblock for artificial intelligence, is impossible for simple automation to cross. There are simply too many rules and nuances that govern the relationship between what a machine sees and what a human understands, and hard-coded rule-based approaches to making this conceptual leap will always be inadequate. Machine learning, though it does not currently perform anywhere near a real human, is the only realistic option humans have to alleviate the work of description for collections at scale.

C. Describing "difficult" material

Beyond alleviating exponentially-growing workloads, machine learning can also provide a method of processing sensitive or challenging materials. Such material is particularly present in human rights archives that handle videos depicting graphic human rights violations.[158] In another sense, archives may find material in other languages challenging to process; machine learning can allow mass classification of video through visual analysis, regardless of language, because of the common vocabularies of visual language. For example, shot boundaries in news broadcasts are identified in the same manner, whether in English or other languages.[159]

---

[158] Aronson.
[159] Smeaton et al, "Video shot boundary detection: Seven years of TRECVID activity," 411-412.

Machine learning can also provide translation of audio and textual content. For example, the project Multilingual Access to Large Spoken Archives (MALACH), at the University of Maryland, worked to make video Holocaust testimonies from the Shoah Visual History Foundation available in English, by creating a pipeline of automatic speech recognition, machine translation, and natural language processing.[160] While machine translation, like speech-to-text transcription, still requires human oversight and editing, such efforts could open up collections to multilingual researchers in a way that archives currently cannot.

D. Answering computational questions

In implementing automation, AI, and any other machine-based description methods, it is important to understand what sorts of questions they can answer. By making use of the structured and pattern-oriented characteristics of digital or digitized materials, researchers can shed a new light on archival materials and library collections at scale. The "Collections as Data" effort, led by Thomas Padilla at the University of Nevada–Las Vegas and funded by the Institute of Museum and Library Services, is a strategic partnership meant to advance the use of collections as data—that is, treating collection data and metadata as *computational* information as well as qualitative.[161] This tactic is key to the field of digital humanities, which lies at the intersection between computing and humanities disciplines and often involves data analysis methods such as mapping, text analysis, and visualizations. [162] Computational analysis of humanities materials—impossible just years ago—has been urged along

[160] "MALACH: Multilingual Access to Large Spoken Archives" (project site), University of Maryland Institute for Advanced Computer Studies, accessed 5 May 2019, https://malach.umiacs.umd.edu/
[161] "Always Already Computational: Collections as Data," *Always Already Computational,* 2018, https://collectionsasdata.github.io/
[162] "Digital Humanities," *Stanford Humanities Center,* Stanford University, accessed 5 May 2019, http://shc.stanford.edu/digital-humanities

by large international initiatives such as the "Digging into Data" challenge (sponsored by the NEH, National Science Foundation, and other national funders in Canada and the United Kingdom).[163]

Such scholarship is untraditional, even disruptive, and relies heavily on collaboration and resource sharing. The first round of "Digging into Data," in 2009, yielded a report from the Council on Library and Information Resources concluding that four distinct kinds of expertise are required for a successful project: subject expertise, analytical expertise, data expertise, and project management expertise.[164] Archivists occupy a middle ground, as the profession's everyday duties requires work in each these areas of expertise, but one person cannot be an expert in each; collaboration, therefore, cannot be glossed over.

Though machines are good at answering computational questions, humans do less well at formulating these questions, partly because they represent such a departure from how humans tend to think about humanities materials. A major part of the HiPSTAS initiative was a set of workshops introducing participant archivists, researchers, and other users to computational analysis; throughout the workshop, participants refined theoretical and vague goals into more specific tasks that engaged with the ARLO technology, but did not exceed it. (For example, one participant began with a hope to perform batch processing, and left with the goal of finding specific audio features such as needle drops to perform batch segmentation.)[165] These thought exercises to refine queries into quantitative puzzles are important to perform before adopting machine learning in the hopes of answering questions that machines cannot answer.

---

[163] Christa Williford, Charles J. Henry, and Amy Friedlander, *One culture: Computationally intensive research in the humanities and social sciences: A report on the experiences of first respondents to the digging into data challenge,* Council on Library and Information Resources, 2012, 8-9.
[164] Williford et al, 2.
[165] Clement et al, "High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities," 6.

3. Practical limitations of machine learning

It is easy to get excited about machine learning as an answer to some of the thorniest problems archivists face. But there are many arenas in which machine learning is inadequate, or even less effective than simple automation. For example, authentication is a goal best accomplished with simple checksums, and fixity and file compliance with standards are easy to measure with automated tools. However, machine learning remains theoretically capable of some tasks that automation is not.

Machine learning also comes with a host of foundational ethical concerns, which will be explored in detail in section IV; the below points examine the extent of its practical ability to accomplish the tasks for which it is best suited.

A. State of the tools

The most obvious limitation of machine learning is its actual effectiveness. At the time this thesis is being written, there are imbalances across artificial intelligence; computer vision research and tools, for example, lag far behind those from the natural language processing field.[166] The actual utility of machine learning tools remains debated, and even the most effective tools tend to perform dramatically worse when fed content that is not highly structured, with good audio, and full of recognizable faces (e.g., television broadcasts or some motion pictures).

Basic tools for audiovisual analysis remain the most effective; in 2012, research professor Tinne Tuytelaars at the Katholieke Universiteit Leuven wrote that "in a more diversified, real world setting, only the most basic tools for content-based audiovisual analysis have really proven their value to date (e.g. shot cut detection, indexing based on ASR output, video copy detection, frontal face detection,

[166] Conversation with Lauren Tilton and Taylor Arnold.

or color based similarity search)."[167] Maciej Cegłowski, at the Library of Congress's "Collections as Data" event, called algorithms

> "as a dim-witted but extremely industrious graduate student, whom you don't fully trust. You want a concordance made? An index? You want them to go through ten million photos and find every picture of a horse? Perfect. You want them to draw conclusions on gender based on word use patterns? Or infer social relationships from census data? Now you need some adult supervision in the room."[168]

Any question that cannot already be answered through analysis of low-level video features or transcriptions may not be answerable without years more of development in the field. Some of the questions that *can* be answered with low-level video features—for example, "how can I find all videos that are predominantly red?"—may not pertain to an archives' mission statement or its users' queries. Indeed, if users are not making this type of query, machine learning may not be the answer to collections access for quite some time. And implementing machine learning before it is ready is not just an issue of quality control, but of trust. If machines generate bad metadata, and if this metadata is not flagged or kept separate, machine learning will dilute good metadata and lower levels of trust in the institution's metadata in general.

B. Quality of the materials

Machine learning also relies heavily on the quality of the materials, and requires significant training for use with analog or low-quality materials. Section III.1.E., "Using collections as datasets," described the potential sources of analytical error for audiovisual materials, and it is worth noting these errors—perfectly acceptable for human comprehension—can render off-the-shelf commercial

---

[167] Tuytelaars, 1.
[168] Maciej Cegłowski, "Deep-Fried Data" (talk, Collections as Data: Stewardship and Use Models to Enhance Access, Library of Congress, Washington, D.C., 27 Sept. 2016), https://idlewords.com/talks/deep_fried_data.htm

products totally useless in archives of analog material. Stanford's Special Collections identified a research project using Allen Ginsberg's audio recordings, in which they hoped to find the points in his recordings at which he turned the recorder on and off. The project was perfect as an application of AI; it identified a simple, bounded sound that a machine could recognize, and that had meaning to human researchers as the beginnings and ends of his dictated poem drafts. However, they had to use an existing commercial tool—Google's speech-to-text service—and found that the results were useless. The tool simply could not analyze analog audio recordings from the 20th century.[169] Some archives may not have the time or technical knowledge to train their own models to deal specifically with analog errors, but many will not.

C. Translating audiovisual materials into text

Text-based search and retrieval, which represent some of the most effective applications of machine learning, suffer from limitations that particularly affect audiovisual materials. As previously discussed, many content management systems do not support time-based access of audiovisual files, and machine-created tags may be confusing or misleading to users who cannot match the tag to the video content. Tuytelaars lists other considerations such as the time required to train corpuses of material on multiple concepts, which limits users to searching from a pre-existing list of concepts. Another limitation is the gap between what machines can understand—even at a relatively high level—and concepts humans might search. It takes a different mindset, and will return many false positives, to search "Soviet Union" or "Gorbachev" or "economics" instead of "Perestroika," an

[169] Conversation with Catherine Nicole Coleman.

academic concept that has little physical representation but that ties together a huge range of materials.[170]

[170] Tuytelaars, 2.

IV. Ethical concerns for archives

Machines are limited by the humans that program them. They can serve as a carrier for bias, ignorance, and oversimplification without appropriate oversight, especially because a machine learning model's inner workings are ultimately opaque even to the people who train it. Machines also have no conception of privacy, nor appreciation of context, and there are limits on how effectively humans can work around these shortcomings.

1. Machine learning as a carrier of bias

Machine learning models can be tricky to evaluate from an ethical standpoint. If it has been created by a machine, based on data, one might expect it to act as a neutral tool to reveal patterns that humans cannot see—and therefore cannot verify. The algorithm itself is indeed entirely a product of its dataset, which in cases of supervised learning is already labeled with the correct answers. But what are the limitations of the dataset? What is the provenance of the data itself? How is the output interpreted? These questions can have serious consequences for the algorithm's current and future output.

A. Biased data

In some cases, the makeup of the training data leads to a tool that is better at processing a certain type of data than others. For example, take a study of commercial image classification software created by IBM, Microsoft, and Face++. Joy Buolamwini and Timnit Gebru demonstrated that these tools performed significantly worse at recognizing images of women (as opposed to men) and

darker-skinned faces (as opposed to lighter).[171] The tools perform this way because of the data with which they were trained. If they had been supplied with more training data of images of women and people of color, then they would be better at recognizing these faces. In this case, it is likely that these companies simply used training datasets that systematically overrepresented images of white men, as many of the largest facial images repositories do.[172] Simply using a training set without auditing the data, therefore, allowed unintentional biases to form the foundation of an algorithm. Even datasets explicitly meant to address disparities can fall short; a 2015 dataset released by the National Institute of Standards and Technology in 2015 that was meant to serve as a geographically diverse benchmark was found to under-represent darker skin tones.[173]

In other cases, algorithms are trained on data that creates misleading relationships. Take a tool that aims to predict the risk of a suspect committing a crime in the next two years. What variables should be used to calculate this risk? Durham, UK, police trained their version of this tool with information including the postal codes of people who had been arrested. When they saw that a certain postal code area had a higher amount of crime, they focused their arrests in that area. Of course, this focus created a feedback loop, and steadily increased the percentage of arrests and scrutiny in that postal code over time.[174] Or take the University of Pittsburgh's attempt to create an algorithm that determined which hospital patients with pneumonia were at the most risk. The model that proved the most effective was also shown to predict that patients with asthma were at less risk than those without. But this learned disparity was due to the fact that hospital staff flagged asthma as a sign of vulnerability and provided more focused care to those patients. If the algorithm had been put in place, it would

[171] As cited in Harper.
[172] Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Conference on Fairness, Accountability and Transparency,* pp. 77-91, in *Proceedings of Machine Learning Research* 81 (2018), 79.
[173] Buolamwini and Gebru, 71-75.
[174] Harper, citing Burgess and Oswald.

have actually directed resources away from patients with asthma, since they were shown to have higher survival rates.[175] Though these particular examples obviously take place outside of an archival context, they illustrate an important question: when algorithms are skewed, how long will it take to recognize and correct them?

B. Biased funding

Machine learning tools and services are disproportionately developed by big tech, and many of the same tools have roots in the defense industry. These industries are not only enormous, but minimally regulated; in fact, they are often asked to regulate themselves. Self-regulation at the expense of income causes an obvious conflict of interest for profit-driven companies.[176] There are also no obvious consequences for violating self-defined technological regulations, because there are no governing boards that can revoke certifications or licenses—in fact, there are no licenses, period. Such incentives and consequences are key to the legal and medical fields, for example, who are understood to be exempted from strict government regulation because of their strict ethical self-governance.[177] Tech regulation, on the other hand, is entirely voluntary and free of consequences, and tech companies hope to keep it that way. Lobbyists claim that government regulation would hinder technology development, and that case law will inevitably emerge to provide some legal governance[178] —though judges are unlikely to understand technology well enough to make completely informed judgements, and the deep pockets of big tech will deeply influence the outcomes.

---

[175] Cliff Kuang, "Can A.I. Be Taught to Explain Itself?," *The New York Times,* 21 Nov. 2017, https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html
[176] Conversation with Enrique Piracés, 21 Jan. 2019.
[177] Amy Brost, "Handling Complex Media in Museums" (lecture, "Handling Complex Media," New York University, New York, NY, 2 Apr. 2019).
[178] "Governing Machine Learning: Exploring the Intersection Between Machine Learning, Law, and Regulation" (white paper), Information Society Project at Yale Law School/Immuta, Sept. 2017, https://law.yale.edu/system/files/area/center/isp/documents/governing_machine_learning_-_final.pdf

These unregulated and profit-driven funding sources are important because they hobble the foundations of the tools that archivists and librarians may use. Profit is a clear driver of some of the most controversial uses of AI today, such as Amazon's contract with Immigration and Customs Enforcement to provide use of its facial recognition tool, Amazon Rekognition, to find and arrest undocumented immigrants.[179] Profit also drives the trickling down of invasive machine learning applications to everyday settings; for example, the ACLU suspects that big box retailers are employing facial recognition techniques on customers without informing them,[180] and police in many states regularly run facial recognition tools against state driver's license and ID databases.[181] Does it violate equality- and privacy-driven library and archival ethics to use tools that undermine those concepts for profit?

Alternative options are few, as these companies are busy consolidating their reach by buying out AI researchers and entire companies. Companies that attempt to "democratize" AI are often acquired by big tech companies who take note of their tools and decide to acquire them; the same companies are famous for identifying and luring away individual researchers from academia, nonprofits, and smaller startups, sometimes attempting to plunder entire departments. Big tech gets away with this tactic both because their monetary resources and their holy grail of machine learning research: huge stockpiles of data, available to employees. This atmosphere of takeover and consolidation of data, talent, and money makes it extraordinarily difficult for individuals and small companies to build tools and conduct research outside the big tech umbrella.[182] This environment is

---

[179] Neema Singh Guliani, "Amazon Met With ICE Officials to Market Its Facial Recognition Product," *Free Future* (blog), American Civil Liberties Union, 24 Oct. 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazon-met-ice-officials-market-its-facial

[180] Jenna Bitar and Jay Stanley, "Are Stores You Shop at Secretly Using Face Recognition on You?," *Free Future* (blog), American Civil Liberties Union, 28 Mar. 2016, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/are-stores-you-shop-secretly-using-face

[181] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle, "The perpetual line-up: Unregulated police face recognition in America," Georgetown Law Center on Privacy & Technology, 18 Oct. 2016, https://www.perpetuallineup.org/

[182] Metz, "Giant Corporations Are Hoarding the World's AI Talent."

particularly damaging within a discipline that already suffers from hype, secrecy, and technological illiteracy—a 2017 *New York Times* article quoted a report that only about 10,000 people in the world are estimated to be capable of "serious artificial intelligence research."[183] True alternatives to big tech tools are even more difficult to find when one considers how many tools are built upon models already trained by this handful of big tech companies, and that training models from scratch is unrealistic.

Machine learning models are wielded by the institutions who can fund and deploy them. When the general public can't wield machine learning in the same way—and especially when the general public doesn't understand what machine learning can and can't do—governments and companies are going to hold all the cards. Anyone trying to use machine learning will be beholden to big tech's training data, models, and priorities in the use of these tools.

C. Biased application

Applications of machine learning often illustrate a deep power imbalance between resource-rich corporations and governments and the general public. One example is that of content takedown algorithms deployed by social media companies such as YouTube and Facebook. These sites rely largely on machine learning algorithms to take down violent, offensive, and illegal content quickly; Facebook also processes flags and reports by machine.[184] However, few companies disclose the principles that guide these algorithms, and even those that do concede that most of their content is removed by machine with no human review—for example, 98% of the violent extremist content removed by YouTube in the last quarter of 2017 was automatically reviewed by machine.[185] In the face

[183] Cade Metz, "Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent," *The New York Times,* 22 Oct. 2017, https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html
[184] Kristina Cooke, "Facebook developing artificial intelligence to flag offensive live videos," *Reuters,* 1 Dec. 2016, https://uk.reuters.com/article/us-facebook-ai-video-idUKKBN13Q52M
[185] Jacob J. Hutt, "Why YouTube Shouldn't Over-Rely on Artificial Intelligence to Police Its Platform," *Free Future* (blog), American Civil Liberties Union, 26 Apr. 2018, https://www.aclu.org/blog/privacy-technology/internet-privacy/why-youtube-shouldnt-over-rely-artificial-intelligence

of such powerful and opaque algorithms, humans have no real recourse to appeal the removal of their video content.

High standards for authenticity and an understanding of the ways in which files can be manipulated are pillars of the information sciences, and have entered the public sphere. Yvonne Ng, Senior Archivist at WITNESS, works with people documenting human rights activism and offenses on video, and has seen the public's expectations of what it takes to demonstrate veracity increase over the last several years as digital video has become common. The idea that "seeing is believing" has always been an oversimplification that does not take into account basic questions (what lies just outside the camera frame? Who is holding the camera?), but the common standards for evidentiary material now often go beyond the images and sound on a video to the presence and reliability of other metadata.[186] Problems arise when evidence is invalidated because people do not have the resources to use the latest tools. This split is starkest in the human rights world, where video is evidentiary in both the court of public opinion and actual court cases. Some of the people who rely most on video content can be shut out of the conversation because their own skills lag behind those of the upper class, the Western world, or large corporations.[187]

Machine learning algorithms have also been applied to automate basic social resources and interactions, having become broadly used to hire, fire, incarcerate, and medically treat people.[188] The wide-ranging implications and ethics of these decisions are the purview of a wide range of researchers and organizations studying algorithmic decision-making and algorithmic fairness, accountability, and transparency.[189] It is no small point that archives and libraries do not predicate such major social

[186] Interview with Yvonne Ng, 17 Jan. 2019.
[187] Interview with Yvonne Ng.
[188] Buolamwini and Gebru, 77.
[189] "FATML Research Centers/People" (collaborative spreadsheet), last updated 26 Mar. 2019, https://docs.google.com/spreadsheets/d/1nacNWHfq1B6SrOC_Dyd-R908qsANeTACG6FibLl9Jls/edit#gid=1530815081

interventions upon the outcome of their own models. However, the application of machine learning in archives still have implications for the materials that can be made accessible. For example, an archives may find that using a voice-to-text product with audio of regional accents creates transcripts riddled with errors, but highly produced media will yield much more correct transcripts. It would not be unusual for an archives strapped for time and staff to have to leave the transcripts as-is—but this choice would make the highly produced media much more discoverable. Or perhaps the archives will try to use machine learning to help label the same oral histories based on whether a woman or a man is speaking. What happens when a voice cannot be sorted into the default conceptions of male and female voices? When the algorithm gets it wrong, either for mislabeling or for a sheer lack of vocabulary to describe non-binary and other voices? These voices are not made discoverable, or even silenced.

2. Concerns for archives

By design, the machine learning training process involves major opportunities for risk. Some of the most basic hazards have to do with the robustness and accuracy of the algorithms, which rely on the data supplied during the training process. Other risks arise from the applications of these algorithms in the outside world, and whom to hold responsible for its effects.

Algorithms have the appearance of objectivity; they take in data and spit back results through a highly technical and inscrutable process. They are, however, supplied with data and employed in society by humans, who remain subject to human error and bias. This truth is most obvious in hard-coded algorithms, such as programs that are coded by humans and executed by machine. If the code fails, the human is the one at fault, and the one responsible for finding the bug and fixing it.

Machine learning introduces the idea that the machine is learning straight from the data. On the surface, this seems like an improvement, and in many ways it can be. Removing the human from the coding process means the machine will reach a solution on its own, without humans intervening and potentially injecting bugs and issues. Because of this process, machine learning algorithms are often thought of as neutral; if humans are removed from the production of the final model, and if machines are not governed by the same assumptions and prejudices as humans, bias will not be embedded in the machine's interpretation of this data. But bias can enter the machine learning process at many steps. Algorithms are no better or worse than the data they receive, the ways they are manipulated, and the ends to which they are used, all of which are subject to human oversight and all of which are likely to reinforce the status quo.

The speciousness of the idea that algorithms are "neutral" has much in common with the speciousness of the idea that libraries and archives are "neutral." Both algorithms and libraries are created and run by humans, and are subject to the assumptions and selections of the people who execute them. These biases may manifest in conscious or unconscious prejudice on the part of the individual, which quickly manifests as systemic oppression in a profession—such as librarianship—that skews 85% white.[190]

A. Preserving context and content

Processing by machine is similar to a ranked form of information retrieval. Machines flag content based on frequency of a keyword or presence of a speaker, and turns users' and archivists'

[190] Chris Bourg, "The unbearable whiteness of librarianship," *Feral Librarian* (blog), 3 Mar. 2014, https://chrisbourg.wordpress.com/2014/03/03/the-unbearable-whiteness-of-librarianship/

attention to those files at the expense of other content. This limitation is important to overcome in both processing and access; if the archivist or the researcher is satisfied with the results returned by the machine, they will forfeit their own understandings of the collection in service of convenience.

Using collections as data strips them of context, as does interpreting them with tools trained on other collections. Any conclusions a machine can come to will never be adequately contextualized; the machine cannot understand the specific circumstances of the material during the act of analyzing, and cannot know what to look for to provide that context. A lack of context is dire, as it makes the communities represented in the materials feel unwelcome and unsafe; weakens researcher conclusions; and even strips records of their evidentiary qualities.

B. Safeguarding privacy

Concerns raised by collections used for machine learning are reminiscent of the concerns raised by the ability to access collections online.[191] Donors could scarcely have imagined that their materials could be used as computational data and used at scale to perform analysis, nor could their donor agreements necessarily be interpreted to give consent to this use. Privacy issues are quite separate from copyright issues and cannot be governed by the same rights and restrictions as copyright imposes. To illustrate this point, Creative Commons (CC) has added a section to its frequently asked questions on the use of CC-licensed works in artificial intelligence software, clarifying for users that CC licenses are not "universal policy tools" and cannot limit reuse of content unless the license terms are explicitly

[191] "Well-intentioned practice for putting digitized collections of unpublished materials online," OCLC Research, OCLC, rev. 28 May 2010, https://www.oclc.org/content/dam/research/activities/rights/practice.pdf

violated (for example, commercial distribution of software made using images under non-commercial licenses).[192]

Privacy concerns also apply to the people represented by the collection. Facial recognition, for example, is an obvious invasion of privacy, one more pertinent for archival collections that represent private individuals (community archives, home movies) than those of public figures (broadcast television, motion pictures). Facial recognition may be a means to a thoughtful end; for example, one of the goals of the Teenie Harris project was to use facial recognition to identify people who appeared in multiple photographs across the more than 80,000-photograph archive of Charles "Teenie" Harris, and present these photos to oral history interviewees in the hopes of being able to name them.[193] However, facial recognition have different privacy implications for disadvantaged groups—because the same tool has difficulty identifying Black faces, only about 5% of the faces identified across multiple photos appeared to be actual matches.[194] This low success rate becomes a privacy issue because mis-identifying subjects is misleading at best (it leads researchers in the wrong direction) and harmful at worst (it may associate people with environments and events that they had nothing to do with).

Privacy concerns also emerge in considering the datasets used to train machine learning tools. The use of images scraped from Google Image Search and public social media posts, without proper licensing or the creator's or subject's consent, is common in both commercial companies and academia.[195] Cases where companies use Creative Commons-licensed or noncommercial datasets seem

---

[192] "Frequently Asked Questions: Artificial intelligence and CC licenses," *Creative Commons,* Creative Commons, 3 May 2019, https://creativecommons.org/faq/#artificial-intelligence-and-cc-licenses

[193] "CMU Wins NEH Grant for Advanced Computer Analysis of Teenie Harris Archive" (press release), Carnegie Mellon University School of Art, 21 Aug. 2017, http://www.art.cmu.edu/news/faculty/teenie-harris-grant/

[194] "Teenie Week of Play" (GitHub repository), Carnegie Museum of Art, last updated Jan. 2019, https://github.com/cmoa/teenie-week-of-play

[195] Rachel Metz, "If your image is online, it might be training facial-recognition AI," *CNN Business,* 19 Apr. 2019, https://www.cnn.com/2019/04/19/tech/ai-facial-recognition/index.html

more clear cut, but even these cases raise questions, as the creator (like a donor) could not have anticipated use of their data in this way. In fact, the creation of commercial tools using any dataset created without consent violates privacy of both creator and subject—whether of faces, objects, audio recordings, or text. One former CEO of a facial recognition company described the use of algorithms developed on these datasets as "the money laundering of facial recognition…laundering the IP and privacy rights out of the faces."[196] Researchers object that it would have been prohibitive to their research to build up such large datasets, and without scraping images, artificial intelligence would not be as far along as it is. One defense they use is that academic work is noncommercial—but academic work is often incorporated into commercial tools.[197] Is it appropriate for libraries and archives to use these tools, built on a foundation of privacy violation and lack of consent, to a noncommercial end?

Though it would be prohibitively labor-intensive to obtain consent from every third party represented in an archival collection—and in many cases impossible due to outdated contact information and death—the at-scale processing that machine learning makes possible removes the archivist even further from an understanding of the collection's areas of risk, confidentiality, and cultural sensitivity. Privacy in the archives relies on the archivist's oversight of the content.[198] Describing collections with machines is only a first step towards responsible stewardship of materials.

C. Transparency

Machine learning techniques defy traditional understandings of transparency. Human decisions can be documented and justified; hard-coded automation can be followed from step to step

[196] Brian Brackeen, quoted in Solon.
[197] Solon.
[198] Ellen LeClere, "As Libraries and Archives Digitize, Implications for Maintaining Individual Privacy," *MediaShift*, May 2016, http://mediashift.org/2016/05/as-libraries-and-archives-digitize-implications-for-maintaining-individual-privacy/

through its entire process. But AI and machine learning are largely unable to explain their own processes, because they can be understood by neither machine nor human. On some level, if machine learning could be explained, there would be no need for it. On another level, machine learning relies heavily on datasets and models whose origins are guarded as commercial property. For these reasons, open source machine learning does not inherently offer full transparency—a cornerstone of the open source credo.[199]

The issue of explainability has caused tension in the adoption of regulations such as the European Union's General Data Protection Regulation, which requires that "any decision made by a machine be readily explainable"—a requirement that most machine learning products are unlikely to meet. There are ways to break down a machine learning model into manageable chunks. For example, researchers may study the cause and effect of individual variables within a model, shedding light on the factors that affect the outcomes of the model at large. But an approach that explains one model will not be able to explain another, because other details and technicalities matter. And neural networks in particular defy explainability, though (paradoxically) researchers are investigating ways to build neural networks that can explain other neural networks.[200]

Even those who work with ML experience serious difficulties in reproducing their results from model to model. The nature of training creates many points at which divergences can creep in, and frameworks and techniques change so quickly that reproduction even a few years later would be nigh impossible. This reproducibility problem makes it difficult to know if a new model has improved on

[199] Kuang.
[200] Kuang.

(or even correctly executed) published models. It also makes it even more difficult to provide transparency.[201]

Transparency is not a perk, but a key to the responsible adoption of machine learning solutions. As archives consultancy AVP puts it, automated and AI solutions

> "must be transparent to support institutions' roles as stewards of collections they are charged with preserving, and to ensure the data they are producing is authentic and trustworthy to the greatest extent possible. Such transparency would not necessarily benefit a commercial entity, as it would expose the core of the system's value. Institutions that value—and indeed trade—on openness, trust, and neutrality, would not have access to the inner workings of the commercial system, making it difficult to fulfill their missions. This would effectively minimize the returns on their own human and financial investments and metadata quality, and the trust in its authenticity over time would suffer."[202]

An opaque solution degrades the core functions of libraries and archives. So much stock is put in the provenance and quality of both data and metadata that any blow to their trustworthiness is devastating. It is crucial to find ways to hold these tools accountable despite their lack of incentive to be open.

D. Artificial limitations on knowledge and discovery

While algorithmic bias can reinforce existing inequalities, the adoption of machine learning poses a much more basic threat to egalitarianism. Artificial intelligence ability is less crucial than literacy in this arena. The inability to see the limitations of algorithms makes them insidious, not just in the results they yield but in the scaffolding they create. Bad algorithms do not just cause bad

[201] Pete Warden, "The Machine Learning Reproducibility Crisis," *Pete Warden's Blog*, 19 Mar. 2018, https://petewarden.com/2018/03/19/the-machine-learning-reproducibility-crisis/

[202] Jon W. Dunn, Juliet L. Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf, "Audiovisual Metadata Platform (AMP) Planning Project: Progress Report & Next Steps," AVP/Indiana University/University of Texas–Austin, Mar. 2018, https://www.weareavp.com/wp-content/uploads/2019/03/AMP-planning-project-report-2018-03-27-1.pdf

decisions, but they change the landscape of decision-making entirely. A bad search algorithm doesn't just push a conspiracy theory website higher in the search results for "Sandy Hook"—it might simultaneously push legitimate news sources down far enough that few people will ever find them. Or it automatically takes down videos without ever letting them emerge onto the landscape of YouTube.

Models are also limited by what they know. A model trained on just one thing will interpret what it sees through the lens of what it has already learned. For example, if a model has been trained to "see" using only images of the ocean, it interpret *all* images it sees as an ocean landscape.[203] There is no real ability for machines to say "I don't know" when faced with an object; machines are always trying to fit what they see into their known world. Training a machine to be able to sort items into an "unknown" class requires showing them images of the "unknown" objects—a task that is difficult to accomplish without feeding the machine images of everything else on earth.[204] In this way, however, machines actually mimic human discovery and interpretation, which is itself limited by prior experiences, knowledge, and worldviews.

[203] Memo Akten, "Learning to See," *Memo.tv Portfolio,* 2017, http://www.memo.tv/portfolio/learning-to-see/
[204] Pete Warden, "What Image Classifiers Can Do About Unknown Objects," *Pete Warden's Blog*, 6 Jul. 2018, https://petewarden.com/2018/07/06/what-image-classifiers-can-do-about-unknown-objects/

V. Frameworks for practice

Archivists and librarians are in an unusual position to respond to the challenges posed by artificial intelligence and machine learning, many of which are not new concepts for a professional community concerned with responsible and transparent access to information. Machine learning methods for handling information and data are subject to bias, barely regulated, and poorly understood. In many ways, machine learning itself is a natural outgrowth and fit of the information sciences. AI is a interdisciplinary field with strong roots in computing, cognitive science, philosophy, and commercial research, among other subjects. In its basis and interdisciplinary structure, it shares many characteristics with the information sciences even as it falls outside of its traditional curriculum.[205]

Frameworks for implementation, collaboration, inspiration, and education must emerge from this history, training, and ethics. Chris Bourg, Director of Libraries at the Massachusetts Institute of Technology, has framed the implications of AI for libraries as a series of questions, asking how libraries can harness AI while regulating it according to professional ethics:

> "1. What role can libraries play in making sure we don't summon the demon; or at least that we have the tools to control or tame the demon?
> 2. How might we leverage AI in support of our missions? How might AI help us do some of our work better?
> 3. How might we support AI and machine learning in ways that are consistent with and natural evolutions of the long-standing missions and functions of libraries as sources of information and the tools, resources, expertise to use that information?"[206]

The following section attempts to answer the questions Bourg raises.

---

[205] Harper.
[206] Chris Bourg, "What happens to libraries and librarians when machines can read all the books?," *Feral Librarian* (blog), 16 Mar. 2017, https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/

1. Auditing tools

Not every algorithm can serve everyone's needs, and in many cases biases can be corrected for by an alert archivist. But it is necessary to consider what materials can actually be made available by machine learning without human mediation, and what this means for archival representation. If an algorithm can only reliably transcribe television anchors' accents and not those of the field segment interviewees; if it consistently recognizes white faces and has trouble "seeing" black ones, and the archivist does not have time to go back through hours of video footage to check for mistakes; if it suggests "homemaker" as a related keyword for "women" but "lawyer" or "banker" for "men" because of the status quo relationships in the closed captioning it was fed; the content of the materials it processes will reinforce current perceptions of societal defaults.

Because machine learning models are opaque even to the humans that train them, it can be difficult to audit an algorithm and avoid skewed assumptions and mistakes that humans would not make. While the models and training data used by these tools constitute a large part of their trustworthiness, the infrastructure that comes with them—their developers and corporate situation, their end user license agreements, and their storage protocols—are all qualities that archives should interrogate. Not all tools were formulated with archival values in mind, a fact that can come through in the terms and conditions of use.

Before implementing a machine learning tool or service, whether proprietary or open, archivists and librarians should interrogate the security and privacy of their data. This conversation may contain the following questions, many of which echo similar conversations about implementing digital storage and digital asset management systems.

Can this task be accomplished through automation?

In considering machine learning tools for application, it is worth considering machine learning's ethical shortcomings in comparison to automation. Step-by-step programming is inherently understandable and editable, which makes automation better at providing transparency, accountability, and the ability to change. If a task can be accomplished through automation, these attributes should weigh heavily in deciding between automation and machine learning; it is far easier to reconcile the former with core archival values than the latter.

Does this service require me to upload data (to perform facial recognition, to produce transcripts from audio files, et cetera)? How much data does it require? How long and under what conditions will this service store my data?

Sharing and storing data with third parties, especially those that are not conceived as digital repositories, may run counter to original donor agreements or local guidelines on privacy and security. If an archive is using MLaaS tools, the data they hope to analyze must be stored in the cloud, which (depending on the terms of service) may violate local policies or restrictions; for example, government archives may be subject to strict privacy and security regulations.

Will my data be used to train this tool? If my data is used to train this tool, will it be used in a private customized model, or will the data be used to train the generic tool? Is my data anonymized in any way?

Some companies offer the opportunity to customize a model that can be used by the user alone; others incorporate the customized model's training back into their generic model. This use may

violate an archives' policies on privacy as well as intellectual property. Some companies, such as Apple, say that they anonymize data—for instance, removing user IDs and encrypting transmissions of data to the cloud.[207] It is up to the archive to decide at what point the benefits of custom training outweigh the privacy considerations. An open source machine learning limitation is subject to the same questions, but is not ruled by commercial interests.

What is the fee model for this tool or service?

User agreements are often subscription-based, and may charge per run or processing event, per the amount of data used, or a flat fee per month. Charging for the amount of data (taking into account the company's data requirements) is a structure worth noting for audiovisual archives, where audio and particularly video takes up far larger amounts of storage than digital manuscript materials.

What was the development and modeling process for this tool? What libraries and datasets did it draw from?

Though this information may be considered proprietary, it is key to the role of being a responsible steward of a collection. Being able to audit the building blocks of a tool provides at least some measure of accountability and provenance in a process that is largely opaque, and allows the institution to trust its own tools and metadata.

What is the tool's license? If the tool is "open," does it incorporate any code from proprietary sources that may not be reused?

[207] Mark Sullivan, "Apple Explains How It's Making Siri Smart Without Endangering User Privacy," *Fast Company,* 11 Sept. 2017, https://www.fastcompany.com/40443055/apple-explains-how-its-making-siri-smart-without-endangering-user-privacy

Open-source projects are often distinguished as driven by public, rather than private. However, the iterative nature of machine learning—the fact that it improves (and is debugged) with more training and more data—means that opening up machine learning libraries for public contributions can potentially aid companies more than they undercut their competitiveness. Perhaps in this spirit, Google and Facebook have both open-sourced machine learning libraries for public use.[208] Nor does openness in one aspect imply openness in all aspects. For example, Google has explicitly open-sourced TensorFlow, its machine learning library, but it owns much of the other code developed by Google employees—meaning projects incorporating code from Google employees may be partly the intellectual property of Google.[209]

Where does the funding for this company come from? Who are its other clients? Who is the intended audience for this tool?

The ethical considerations of using tools created by military or Silicon Valley companies have already been discussed at length. These questions do not preclude the possibility that open or nonprofit tools can still be prey for larger corporations and money-making ventures. For example, take Pop Up Archive, an audio transcription service whose clients included audiovisual archives such as the Studs Terkel Radio Archive and WGBH—a sure sign that they were invested in cultural heritage institutions. Pop Up Archive was funded by the National Endowment for the Humanities, the Knight Foundation, and the Institute of Museum and Library Services; however, they shut down to clients once they were bought by Apple in 2017.[210] Such a development is difficult to predict in advance but worth keeping in mind.

[208] "Google leads in the race to dominate artificial intelligence."
[209] Conversation with Winnie Schwaid-Lindner, 1 Feb. 2019.
[210] Howard.

Whether or not the library and archives world is one of the intended audiences of a tool often—though not always—correlates to the mission statement and driving forces of the company that creates the tool. Companies intent on profit are less likely to account for the unique needs of audiovisual archives. Grant-funded groups or individuals in academia or archives are far more likely to make their tool freely available and gear it towards the concerns of the information sciences.

Is this tool appropriate for the needs of my collection? What do this tool's results look like for a diverse sample of my collection?

It is important to test out tools on local collections to have clear expectations about what it will be able to do for an archives. Demonstrations done by a vendor will often involve examples that have been vetted and that are known to work; see what the tool will do with representative idiosyncratic materials from the collections.

Can I understand and explain the limitations of this tool to others? Does this tool report error rates and levels of confidence?

It is important for archivists to have a sense of how effective a tool is, what caveats to apply to its results, and how to communicate its limitations to others. This process is made easier if the tool itself self-reports using confidence intervals, or percentages that represents how sure the machine is of its classification.[211]

[211] Russell and Norvig, 761.

2. Collaboration

Many of the issues that plague these tools within the context of the information sciences—their inability to understand and support for analog recordings, their lack of privacy safeguards, and their limitations on large amounts of data, among others—are solvable with collaboration between libraries and archives. Funding plays a huge role in this environment. Libraries and archives may not offer a commercial incentive for corporations to accommodate their requests and needs, even though commercial solutions are the most cost- and labor-effective; grant-funded projects that take these needs into account require huge investments of time by a small team, and are usually over within a few years at most. Large-scale collaboration is a clear path forward. Nonprofit institutions—academic and governmental institutions in particular—need to join together to develop a number of key resources. One resource is datasets that represent typical collection items and formats, that have been gauged to be appropriate for training use and excised of sensitive content. Another resource is tools that can train on these datasets and yield transcripts, tags, and other results for the same archival materials. Yet another resource would be a set of best practices for the library and archives community.

Libraries and archives also need to draw from other communities that are thinking about ethics, best practices, and regulations. A long list of researchers and academic organizations, ranging from the Harvard Berkman Center to the UCLA professor Safiya Noble, are engaging with moral critiques and ethical governance of AI. Resources from collaborative spreadsheets[212] to class syllabi[213] to councils[214] to task forces[215] to antitrust initiatives[216] provide position papers, best practices, and

[212] "FATML Research Centers/People."
[213] Joi Ito and Jonathan Zittrain, "The Ethics and Governance of Artificial Intelligence" (syllabus), *Massachusetts Institute of Technology Media Lab,* spring 2018, https://www.media.mit.edu/courses/the-ethics-and-governance-of-artificial-intelligence/
[214] "Council for Big Data, Ethics, and Society," *Council for Big Data, Ethics, and Society,* National Science Foundation, accessed 6 May 2019, https://bdes.datasociety.net/council-members/

analysis upon which the library and archives community can draw—without having to replicate the same work.

Libraries and archives can even point to these resources as models. Though the US government and Silicon Valley are minimally regulated in general, and AI in particular is so new that it is poorly understood by regulators, some guidelines do exist. Libraries and archives can comply with existing regulations, such as the EU's General Data Protection Regulation (GDPR), which requires that organizations be able to explain their algorithmic decisions.[217]

3. Machine learning literacy

Machine learning literacy applies to archivists as well as users of archives. In order to leverage machine learning to its greatest efficacy, those who hope to describe and access collections must first understand how machine learning works, and thereby understand its limitations. Machine learning is useless if there it turns out there is not a demand for its services; if an archives finds it does not ultimately need transcripts, but narrative reference guides, a machine learning solution will not be appropriate. In other cases, automation solutions will be more appropriate. If a collection has already been cataloged, it is a far lighter use of resources to spend some time on extracting that metadata than entirely re-processing the collection.

In many cases, the very concepts of AI must be demystified. Many librarians and archivists see AI as the domain of technologists and are not given space to explore it within their day-to-day roles; such an environment shuts out the people whose understanding of collections and context is crucial to

[215] "New York City Automated Decision Systems Task Force," *Website of the City of New York,* City of New York, accessed 6 May 2019, https://www1.nyc.gov/site/adstaskforce/index.page
[216] Michaela Ross, "Artificial Intelligence Pushes the Antitrust Envelope," *Bloomberg Law*, 28 Apr. 2017, https://perma.cc/3JV5-P9VK
[217] Kuang.

the success of any AI project. Nicole Coleman, the Digital Research Architect at Stanford University Libraries and Research Director for the Humanities + Design Research Lab, describes the crux of her work as raising the level of AI literacy across the institution—a simply stated goal, but one which must change prevailing conceptions of librarianship along the way.[218]

Machine learning illiteracy presents the potential for extremely wide gulfs between the general public and technologists. In its current state, machine learning is new enough to present challenges and scenarios that are unfamiliar to the general public. It is also new enough that most applications are deployed at scale by big corporations, who can afford to hire the talent, obtain the data, and pay for the computing power necessary to generate algorithms.[219] The average person, therefore, not only is unlikely to possess enough technological literacy to train their own machine learning model, but is not sure how these models could be relevant to their own life. Those who manage to teach themselves how to leverage machine learning are unlikely to do so effectively enough to counteract the efforts of large companies and governments.

Machine learning literacy, however, also involves skepticism. The ethics of machine learning are not a sidebar to the technology, but the key to its responsible and contextual use. Even the ability to program a machine to perform description tasks is counterproductive if one does not understand the limitations and biases displayed in the results.

[218] Conversation with Catherine Nicole Coleman, 30 Apr. 2019.
[219] Cade Metz, "Giant Corporations Are Hoarding the World's AI Talent," *Wired,* 17 Nov. 2016, https://www.wired.com/2016/11/giant-corporations-hoarding-worlds-ai-talent/

VI. Conclusion

Chris Bourg posits the idea of the algorithm as "a new kind of patron." Just as librarians are crucial intermediaries for patrons in finding data that will yield useful results, they are intermediaries between algorithms and information.[220] But maximizing data for human discovery and algorithmic discovery may mean very different things. Machines think differently, see differently, hear differently, and understand differently; archivists and users need to begin thinking as machines do, to make full use of their capabilities.

The adoption of new techniques for digital processing also require reckoning with the effects of bias, which manifest themselves at many levels. The capitalist themes that characterize the majority of machine learning spending (defense, service, advertising) naturally slew tools that benefit corporations. On the side of the coin, the spending decisions made by low-funded cultural heritage institutions have cascading effects on their mission; for example, cutting the processing budget gives donors with means an easy way to jump to the head of the queue. It has always been important for archives and libraries not to relax in an assumed "neutrality," but to stand on guard against human bias. Opaque and ill-trained models are simply a new frontier.

Such tools do not replace human judgment in decisions that have a direct impact on society and human lives, nor do these algorithms attempt to outreach human comprehension. The machine learning methods detailed in this thesis ask machines to accomplish rote, yet complex, classification tasks that otherwise would not be achievable on a human timescale. At the same time, allowing machines the full responsibility of processing and surfacing audiovisual content will inevitably yield a

[220] Bourg, "What happens to libraries and librarians when machines can read all the books?"

sample unrepresentative of curatorial and archival concerns. The ethical concerns of these tools present less of a roadblock than a framework for future algorithmic development.

Machine learning is a clear help to archives hoping to make transcripts available, or segment their materials, or reviewing traumatic materials without constant human burnout. These applications do not suffer from the same deep ethical questions as facial recognition might; in many ways, they just serve as ways to facilitate MPLP. This state of things is fine! Machine learning does not have to solve archival description processes all at once. It needs to start by tangibly making our lives as archivists easier.

In a world where archivists are simply trying to triage the amount of born-digital material up for processing, any amount of automation helps. But until algorithms grow sophisticated enough to encompass the true breadth of archival collections, it will be crucial to remain aware of the imbalances machine learning methods create, and to work to address this imbalance. This work can only be accomplished by humans—meaning that archivists and librarians are not about to be replaced. Algorithms are often trained by the same humans who give their output meaning. It is important to take the role of human seriously and critically, because the machine cannot correct for a human's failings.

Works cited

Interviews

Clancy, Eileen. 5 Mar. 2019.

Coleman, Catherine Nicole. 30 Apr. 2019.

Ng, Yvonne. 17 Jan. 2019.

Piracés, Enrique. 21 Jan. 2019.

Tilton, Lauren, and Taylor Arnold. 16 Apr. 2019.

Automation, AI, and ML in libraries and archives

General resources and overviews

"Always Already Computational: Collections as Data." *Always Already Computational*. 2018. https://collectionsasdata.github.io/

Bourg, Chris. "What happens to libraries and librarians when machines can read all the books?" *Feral Librarian* (blog). 16 Mar. 2017. https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/

Cegłowski, Maciej. "Deep-Fried Data" (talk). *Collections as Data: Stewardship and Use Models to Enhance Access*, Library of Congress, Washington, D.C. 27 Sept. 2016. https://idlewords.com/talks/deep_fried_data.htm

Coleman, Catherine Nicole. "Artificial intelligence and the library of the future, revisited." *Digital Library Blog*, Stanford University Libraries. 3 Nov. 2017. http://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited

Coleman, Catherine Nicole. "Library-Inspired Artificial Intelligence: Discovery, Part 1." *Digital Library Blog*, Stanford University Libraries. 22 Oct. 2018. http://library.stanford.edu/blogs/digital-library-blog/2018/10/library-inspired-artificial-intelligence-discovery-part-1

Harper, Charlie. "Machine Learning and the Library or: How I Learned to Stop Worrying and Love My Robot Overlords." *Code4Lib Journal* 41 (Aug. 2018). https://journal.code4lib.org/articles/13671

Nowviskie, Bethany. "Reconstitute the World." Bethany Nowviskie (personal website). 12 June 2018. http://nowviskie.org/2018/reconstitute-the-world/

Case studies in archives and libraries

Automation, artificial intelligence, and machine learning led by librarians and archivists.

"BitCurator NLP." *BitCurator.net*. BitCurator. Accessed 2 May 2019. https://bitcurator.net/bitcurator-nlp/

Boyd, Doug. "OHMS: Enhancing access to oral history for free." *The Oral History Review* 40, no. 1 (2013): 95-106.

Clement, Tanya E., Loretta Auvil, and David Tcheng. "High Performance Sound Technologies for Access and Scholarship" (white paper). 2016.

Clement, Tanya E., David Tcheng, Loretta Auvil, and Tony Borries. "High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities." *Proceedings of the American Society for Information Science and Technology* 51, no. 1 (2014): 1-10.

Clement, Tanya. "Introducing the HiPSTAS Audio Toolkit Workflow: Audio Labeling." *High Performance Sound Technologies for Access and Scholarship (HiPSTAS)* (blog), University of Texas at Austin. https://blogs.ischool.utexas.edu/hipstas/2017/08/31/introducing-the-hipstas-audio-toolkit-workflow-audio-labeling/

"CMU Wins NEH Grant for Advanced Computer Analysis of Teenie Harris Archive" (press release). Carnegie Mellon University School of Art. 21 Aug. 2017. http://www.art.cmu.edu/news/faculty/teenie-harris-grant/

Dunn, Jon W., Juliet L. Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf. "Audiovisual Metadata Platform (AMP) Planning Project: Progress Report & Next Steps." AVP/Indiana University/University of Texas–Austin. Mar. 2018. https://www.weareavp.com/wp-content/uploads/2019/03/AMP-planning-project-report-2018-03-27-1.pdf

Flueckiger, Barbara. "A Digital Humanities Approach to Film Colors." *Moving Image: The Journal of the Association of Moving Image Archivists* 17, no. 2 (2017). 71-94.

"High-Performance Sound Technologies for Access and Scholarship" (GitHub user page). HiPSTAS at University of Texas at Austin. Accessed 4 May 2019. https://github.com/hipstas

Howard, Jennifer. "Pop Up Archive Filled a Need for Audio Archiving, and Apple Noticed." *Humanities: The Magazine of the National Endowment for the Humanities* 38, no. 4 (Fall 2017). National Endowment for the Humanities. https://www.neh.gov/humanities/2017/fall/feature/pop-archive-filled-need-audio-archiving-and-apple-noticed

Jaquith, Waldo. "How I OCR hundreds of hours of video" (blog post). 10 Feb. 2011. https://waldo.jaquith.org/blog/2011/02/ocr-video/

"MALACH: Multilingual Access to Large Spoken Archives" (project site). University of Maryland Institute for Advanced Computer Studies. Accessed 5 May 2019. https://malach.umiacs.umd.edu/

Miotto, R., and Orio, N. "Accessing Music Digital Libraries by Combining Semantic Tags and Audio Content." In *Italian Research Conference on Digital Libraries*, 26-37. Berlin: Springer, 2011.

Mosseri, Inbar, and Oran Lang. "Looking to Listen: Audio-Visual Speech Separation." *Google AI Blog*. Google. 11 Apr. 2018. https://ai.googleblog.com/2018/04/looking-to-listen-audio-visual-speech.html

"Neural Neighbors: Capturing Image Similarity." *Digital Humanities Laboratory.* Yale University Library. Accessed 3 May 2019. http://dhlab.yale.edu/projects/neural_neighbors.html

"Oral History Metadata Synchronizer." *Oral History Metadata Synchronizer.* Louie B. Nunn Center for Oral History at the University of Kentucky Libraries. 2019. http://www.oralhistoryonline.org/

Pustejovsky, James, Nancy Ide, Marc Verhagen, and Keith Suderman. "Enhancing Access to Media Collections and Archives Using Computational Linguistic Tools." In *CDH@TLT*, pp. 19-28. 2017.

"Teenie Week of Play" (GitHub repository). Carnegie Museum of Art. Last updated Jan. 2019. https://github.com/cmoa/teenie-week-of-play

Tilton, Lauren, and Taylor Arnold. "NEH Grant Narrative: Distant Viewing." *Distant Viewing.* University of Richmond/National Endowment for the Humanities. 2018. https://distantviewing.org/pdf/neh_grant_narrative.pdf

Weaver, Andrew. "Adventures in Perceptual Hashing" (blog post). AAPB NDSR blog. National Digital Stewardship Residency. 20 Apr. 2017. https://ndsr.americanarchive.org/2017/04/20/adventures-in-perceptual-hashing/

Webb, Sharon, Chris Kiefer, Ben Jackson, James Baker, and Alice Eldridge. "Mining oral history collections using music information retrieval methods." *Music Reference Services Quarterly* 20, no. 3-4 (2017): 168-183.

Williford, Christa, Charles J. Henry, and Amy Friedlander. *One culture: Computationally intensive research in the humanities and social sciences: A report on the experiences of first respondents to the digging into data challenge.* Council on Library and Information Resources, 2012.

General resources and background on automation, AI, and ML

Mechanics of AI

3Blue1Brown, "But what *is* a Neural Network? | Deep learning, chapter 1." YouTube video, 19:13. *3Blue1Brown Series* season 3, episode 1. 5 Oct. 2017. https://www.youtube.com/watch?v=aircAruvnKk&feature=youtu.be

Akten, Memo. "Learning to See." *Memo.tv Portfolio*. 2017. http://www.memo.tv/portfolio/learning-to-see/

Apple Computer Vision Machine Learning Team. "An On-device Deep Neural Network for Face Detection." *Apple Machine Learning Journal* 1, no. 7 (Nov. 2017). https://machinelearning.apple.com/2017/11/16/face-detection.html

Dieleman, Sander. "Recommending music on Spotify with deep learning" (blog post). http://benanne.github.io/2014/08/05/spotify-cnns.html

Domingos, Pedro. "A few useful things to know about machine learning." *Communications of the ACM* 55, no. 10 (2012): 78-87.

Grigorev, Max. "Keeping up with AI in 2019." *The Launchpad* (blog). 14 Feb. 2019. https://medium.com/thelaunchpad/what-is-the-next-big-thing-in-ai-and-ml-904a3f3345ef

Hirschberg, Julia, and Christopher D. Manning. "Advances in natural language processing." *Science* 349, no. 6245 (2015): 261-266.

Hof, Robert D. "Deep Learning." *MIT Technology Review.* 23 Apt. 2013. https://www.technologyreview.com/s/513696/deep-learning/

"How to choose algorithms for Azure Machine Learning Studio." *Microsoft Azure Machine Learning Studio Documentation.* 3 Mar. 2019. https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

Huang, Xuedong, James Baker, and Raj Reddy. "A historical perspective of speech recognition." *Communications of the ACM* 57, no. 1 (2014), 94-103.

Kuang, Cliff. "Can A.I. Be Taught to Explain Itself?" *The New York Times.* 21 Nov. 2017. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization: How neural networks build up their understanding of images." *Distill* 2, no. 11 (2017): e7. https://distill.pub/2017/feature-visualization/

Palani, Pradeep. "Understanding Semantic Analysis (And Why This Title is Totally Meta)." *Zeta Global.* Zeta Global blog. 10 Jan. 2018. https://zetaglobal.com/blog-posts/understanding-semantic-analysis-title-totally-meta/

Phillips, Winfred. "Introduction to Natural Language Processing." *Consortium on Cognitive Science Instruction.* Illinois State University. 2006. http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php

Rothmann, Daniel. "What's wrong with CNNs and spectrograms for audio processing?" *Towards Data Science* (blog). Towards Data Science Inc. 25 Mar. 2018. https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd

Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. 3rd ed. Upper Saddle River, New Jersey: Pearson Education Limited, 2010.

Tuytelaars, Tinne. "Content-based analysis for accessing audiovisual archives: alternatives for concept-based indexing and search." In *2012 13th International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1-4. IEEE, 2012.

Warden, Pete. "The Machine Learning Reproducibility Crisis." *Pete Warden's Blog.* 19 Mar. 2018. https://petewarden.com/2018/03/19/the-machine-learning-reproducibility-crisis/

Warden, Pete. "What Image Classifiers Can Do About Unknown Objects." *Pete Warden's Blog.* 6 Jul. 2018. https://petewarden.com/2018/07/06/what-image-classifiers-can-do-about-unknown-objects/

AI and ethics

Aronson, Jay D. "Computer Vision and Machine Learning for Human Rights Video Analysis: Case Studies, Possibilities, Concerns, and Limitations." *Law & Social Inquiry* 43, no. 4 (2018): 1188-1209.

Bitar, Jenna, and Jay Stanley. "Are Stores You Shop at Secretly Using Face Recognition on You?" *Free Future* (blog). American Civil Liberties Union. 28 Mar. 2016. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/are-stores-you-shop-secretly-using-face

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on Fairness, Accountability and Transparency,* pp. 77-91. In *Proceedings of Machine Learning Research* 81 (2018).

Chen, Angela. "Inmates in Finland are training AI as part of prison labor." *The Verge.* 28 Mar. 2019. https://www.theverge.com/2019/3/28/18285572/prison-labor-finland-artificial-intelligence-data-tagging-vainu

"Council for Big Data, Ethics, and Society." *Council for Big Data, Ethics, and Society.* National Science Foundation. Accessed 6 May 2019. https://bdes.datasociety.net/council-members/

"FATML Research Centers/People" (collaborative spreadsheet). Last updated 26 Mar. 2019. https://docs.google.com/spreadsheets/d/1nacNWHfq1B6SrOC_Dyd-R908qsANeTACG6FibLl9Jls/edit#gid=1530815081

"Frequently Asked Questions: Artificial intelligence and CC licenses." *Creative Commons.* Creative Commons. 3 May 2019. https://creativecommons.org/faq/#artificial-intelligence-and-cc-licenses

Garvie, Clare, Alvaro Bedoya, and Jonathan Frankle. "The perpetual line-up: Unregulated police face recognition in America." Georgetown Law Center on Privacy & Technology. 18 Oct. 2016. https://www.perpetuallineup.org/

"Governing Machine Learning: Exploring the Intersection Between Machine Learning, Law, and Regulation" (white paper). Information Society Project at Yale Law School/Immuta. Sept. 2017. https://law.yale.edu/system/files/area/center/isp/documents/governing_machine_learning_-_final.pdf

Guliani, Neema Singh. "Amazon Met With ICE Officials to Market Its Facial Recognition Product." *Free Future* (blog). American Civil Liberties Union. 24 Oct. 2018. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazon-met-ice-officials-market-its-facial

Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. "A data-driven analysis of workers' earnings on Amazon Mechanical Turk." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 449. ACM, 2018.

Hutt, Jacob J. "Why YouTube Shouldn't Over-Rely on Artificial Intelligence to Police Its Platform." *Free Future* (blog). American Civil Liberties Union. 26 Apr. 2018. https://www.aclu.org/blog/privacy-technology/internet-privacy/why-youtube-shouldnt-over-rely-artificial-intelligence

Ito, Joi, and Jonathan Zittrain. "The Ethics and Governance of Artificial Intelligence" (syllabus). *Massachusetts Institute of Technology Media Lab.* Spring 2018. https://www.media.mit.edu/courses/the-ethics-and-governance-of-artificial-intelligence/

Metz, Cade. "Giant Corporations Are Hoarding the World's AI Talent." Wired. 17 Nov. 2016. https://www.wired.com/2016/11/giant-corporations-hoarding-worlds-ai-talent/

Metz, Rachel. "If your image is online, it might be training facial-recognition AI." *CNN Business.* 19 Apr. 2019. https://www.cnn.com/2019/04/19/tech/ai-facial-recognition/index.html

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3, no. 2 (2016): 1-21.

"New York City Automated Decision Systems Task Force." *Website of the City of New York.* City of New York. Accessed 6 May 2019. https://www1.nyc.gov/site/adstaskforce/index.page

Penn, Jonnie. "AI thinks like a corporation—and that's worrying." *Open Future* (blog). *The Economist.* 26 Nov. 2018. https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying

Ross, Michaela. "Artificial Intelligence Pushes the Antitrust Envelope." *Bloomberg Law.* 28 Apr. 2017. https://perma.cc/3JV5-P9VK

Simonite, Tom. "To Make AI Smarter, Humans Perform Oddball Low-Paid Tasks." *Wired.* 9 Feb. 2018, https://www.wired.com/story/behind-artificial-intelligence-lurk-oddball-low-paid-tasks/

Solon, Olivia. "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent." *NBC News.* 12 Mar. 2019. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

Yuan, Li. "How Cheap Labor Drives China's A.I. Ambitions." *The New York Times.* 25 Nov. 2018. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html

AI, business, and politics

"6 Best Practices for Implementing AI in a DAM." *Media Valet* (blog). Mediavalet. Accessed 5 May 2019. https://www.mediavalet.com/blog/implementing-artificial-intelligence-in-dam/

"13 Startups Transcribing Voice to Text Using AI." *Nanalyze.com*. Nanalyze. 29 July 2018. https://www.nanalyze.com/2018/07/voice-to-text-transcribing-ai/

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. "The Obama Administration's Roadmap for AI Policy." *Harvard Business Review.* 21 Dec. 2016. https://hbr.org/2016/12/the-obama-administrations-roadmap-for-ai-policy

"Big Tech In AI: What Amazon, Apple, Google, GE, And Others Are Working On." *CB Insights*. 12 Oct. 2017. https://www.cbinsights.com/research/top-tech-companies-artificial-intelligence-expert-intelligence/

Cooke, Kristina. "Facebook developing artificial intelligence to flag offensive live videos." *Reuters.* 1 Dec. 2016. https://uk.reuters.com/article/us-facebook-ai-video-idUKKBN13Q52M

"Google leads in the race to dominate artificial intelligence." *The Economist* (print). 7 Dec. 2017. https://www.economist.com/business/2017/12/07/google-leads-in-the-race-to-dominate-artificial-intelligence

Harwell, Drew. "Defense Department pledges billions toward artificial intelligence research." *The Switch* (blog). *The Washington Post*. 7 Sept. 2018. https://www.washingtonpost.com/technology/2018/09/07/defense-department-pledges-billions-toward-artificial-intelligence-research/

Jarnow, Jesse. "Transcribing Audio Sucks—So Make The Machines Do It." *Wired*. 26 Apr. 2017. https://www.wired.com/2017/04/trint-multi-voice-transcription/

Kharpal, Arjun. "China wants to be a $150 billion world leader in AI in less than 15 years." *Tech Transformers*. CNBC. 21 Jul. 2017. https://www.cnbc.com/2017/07/21/china-ai-world-leader-by-2030.html

Lee, Timothy B. "The hype around driverless cars came crashing down in 2018." *Ars Technica.* 30 Dec. 2018.

https://arstechnica.com/cars/2018/12/uber-tesla-and-waymo-all-struggled-with-self-driving-in-2018/

Light, Rob. "Machine Learning as a Service (MLaaS)." *G2 Digital Trends* (blog). G2 Crowd. Accessed 5 May 2019. https://blog.g2crowd.com/blog/trends/artificial-intelligence/2018-ai/machine-learning-service-mlaas/

Metz, Cade. "Facebook's AI Is Now Automatically Writing Photo Captions." *Wired.* 5 Apr. 2016. https://www.wired.com/2016/04/facebook-using-ai-write-photo-captions-blind-users/

Metz, Cade. "Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent." *The New York Times.* 22 Oct. 2017. https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html

Metz, Cade. "Trump Signs Executive Order Promoting Artificial Intelligence." *The New York Times.* 11 Feb. 2019. https://www.nytimes.com/2019/02/11/business/ai-artificial-intelligence-trump.html

MSV, Janakiram. "An Executive's Guide to Understanding Cloud-Based Machine Learning Services." *Forbes.com.* 1 Jan. 2019. https://www.forbes.com/sites/janakirammsv/2019/01/01/an-executives-guide-to-understanding-cloud-based-machine-learning-services/#52d5709e3e3e

Su, Jean Baptiste. "Venture Capital Funding For Artificial Intelligence Startups Hit Record High In 2018." *Forbes.com.* 12 Feb. 2019. https://www.forbes.com/sites/jeanbaptiste/2019/02/12/venture-capital-funding-for-artificial-intelligence-startups-hit-record-high-in-2018/#19e6357f41f7

Sullivan, Mark. "Apple Explains How It's Making Siri Smart Without Endangering User Privacy." *Fast Company.* 11 Sept. 2017. https://www.fastcompany.com/40443055/apple-explains-how-its-making-siri-smart-without-endangering-user-privacy

Wilson, Martin. "AI in DAM: The Challenges and Opportunities." *Digital Asset Management News.* Accessed 5 May 2019. https://digitalassetmanagementnews.org/features/ai-in-dam-the-challenges-and-opportunities/

Case studies

Projects and products not led or created by archives and libraries.

"About CMUSphinx." *CMUSphinx Open Source Speech Recognition Toolkit.* CMUSphinx. https://cmusphinx.github.io/wiki/about/

Agnihotri, Lalitha, Kavitha Vallari Devara, Thomas McGee, and Nevenka Dimitrova. "Summarization of video programs based on closed captions." In *Storage and Retrieval for Media Databases* 2001, vol. 4315, pp. 599-608. International Society for Optics and Photonics, 2001.

"Amazon Rekognition." *Amazon Web Services.* Amazon Web Services, Inc. Accessed 1 May 2019. https://aws.amazon.com/rekognition/

"Amazon Transcribe." *Amazon Web Services.* Amazon Web Services, Inc. Accessed 5 May 2019. https://aws.amazon.com/transcribe/

Aronson, Jay D., Shicheng Xu, and Alex Hauptmann. "Video analytics for conflict monitoring and human rights documentation" (technical report). Center for Human Rights Science Technical Report, Carnegie Mellon University (2015).

Chua, Tat-Seng, Shih-Fu Chang, Lekha Chaisorn, and Winston Hsu. "Story boundary detection in large broadcast news video archives: techniques, experience and trends." In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 656-659. ACM, 2004.

Cohen, Patricia. "Digital Keys for Unlocking the Humanities' Riches." *The New York Times.* 16 Nov. 2010. https://www.nytimes.com/2010/11/17/arts/17digital.html?_r=0

Feng, Weijiang, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo. "Audio visual speech recognition with multimodal recurrent neural networks." In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 681-688. IEEE, 2017.

"The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)," ImageNet, Stanford Vision Lab, 2017, http://www.image-net.org/challenges/LSVRC/

"History of the Kaldi project." *Kaldi documentation.* Kaldi. Accessed 2 May 2019. http://kaldi-asr.org/doc/history.html

Lienhart, Rainer W., and Frank Stuber. "Automatic text recognition in digital videos." In *Image and Video Processing* IV, vol. 2666, pp. 180-189. International Society for Optics and Photonics, 1996.

Liu, Zhu, Yao Wang, and Tsuhan Chen. "Audio feature extraction and analysis for scene segmentation and classification." *Journal of VLSI signal processing systems for signal, image and video technology* 20, no. 1-2 (1998): 61-79.

"OpenPose" (GitHub repository). Carnegie Mellon University Perceptual Computing Lab. Accessed 4 May 2019. https://github.com/CMU-Perceptual-Computing-Lab/openpose

Poplin, Ryan, et al. "Predicting cardiovascular risk factors from retinal fundus photographs using deep learning." arXiv preprint arXiv:1708.09843 (2017). Cited in Coleman, "Library-Inspired Artificial Intelligence."

"Pricing." *Trint.com.* Trint. Accessed 5 May 2019. https://trint.com/pricing/

Smeaton, Alan F., Paul Over, and Wessel Kraaij. "High-level feature detection from video in TRECVID: a 5-year retrospective of achievements." In *Multimedia content analysis*, pp. 1-24. Springer, Boston, MA, 2009.

Smeaton, Alan F., Paul Over, and Aiden R. Doherty. "Video shot boundary detection: Seven years of TRECVID activity." *Computer Vision and Image Understanding* 114, no. 4 (2010): 411-418.

Tao, Fei, and Carlos Busso. "Gating neural network for large vocabulary audiovisual speech recognition." *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26, no. 7 (2018): 1286-1298.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "ImageNet large scale visual recognition challenge." *International journal of computer vision* 115, no. 3 (2015): 211-252.

"Train a Model for Custom Speech." *Microsoft Azure documentation.* Microsoft. 1 May 2019. https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-train-model

Ye, Qixiang, and David Doermann. "Text detection and recognition in imagery: A survey." *IEEE transactions on pattern analysis and machine intelligence* 37, no. 7 (2015): 1480-1500.

Other works cited

Abbott, Leala. "The DAM List" (spreadsheet). Created Feb. 2011. https://docs.google.com/spreadsheets/d/1xRwkQVluqtlLVeuLqHtx3EtZeNAye3n_7BwR13GKwm0/edit#gid=0

Arroyo-Ramirez, Elvia. "Invisible Defaults and Perceived Limitations: Processing the Juan Gelman Files" (blog post). 30 Oct. 2016. https://medium.com/on-archivy/invisible-defaults-and-perceived-limitations-processing-the-juan-gelman-files-4187fdd36759

Blewer, Ashley. "The Collection Management System Collection" (collaborative spreadsheet). Created Aug. 2017. https://docs.google.com/spreadsheets/d/1cXOug3qM0pNNeD_wssiVEv9c0W1Y5I1VDTnSPTk7fb4/edit#gid=0

Bourg, Chris. "The unbearable whiteness of librarianship." *Feral Librarian* (blog). 3 Mar. 2014. https://chrisbourg.wordpress.com/2014/03/03/the-unbearable-whiteness-of-librarianship/

Breeding, Marshall. "Library Systems Report 2018." *American Libraries Magazine*, American Libraries Magazine (website). 1 May 2018. https://americanlibrariesmagazine.org/2018/05/01/library-systems-report-2018/

Brost, Amy. "Handling Complex Media in Museums" (lecture). "Handling Complex Media," New York University, New York, NY. 2 Apr. 2019.

"Can ffmpeg extract closed caption data" (Stack Overflow question). Posted by spinon. 3 Jul. 2010. https://stackoverflow.com/questions/3169910/can-ffmpeg-extract-closed-caption-data

Ciocca, Sophie. "How Does Spotify Know You So Well?" (blog post). 10 Oct. 2017. https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe

"Create Accessible Video, Audio and Social Media." *Section508.gov.* U.S. General Services Administration. May 2018. https://www.section508.gov/create/video-social

Conversation with Dave Rice. Oct. 2019.

Conversation with Winnie Schwaid-Lindner. Feb. 1 2019.

"Digital Humanities." *Stanford Humanities Center.* Stanford University. Accessed 5 May 2019. http://shc.stanford.edu/digital-humanities

"File Information Tool Set (FITS)." *Projects at Harvard*. Harvard University. 2019. https://projects.iq.harvard.edu/fits/home

Geraci, Noah. "Programmatic approaches to bias in descriptive metadata" (presentation). *Code4Lib 2019*, San José, CA. Feb. 2019. https://osf.io/9uehx/

Groover, Mikell P. "Automation." *Britannica.com.* Encyclopedia Britannica. 22 Mar. 2019. https://www.britannica.com/technology/automation

"Guideline 1.2: Time-based media." *Web Content Accessibility Guidelines (WCAG) 2.0.* World Wide Web Consortium, eds. Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, et al. 11 Dec. 2008. https://www.w3.org/TR/WCAG20/

"How to Meet WCAG 2 (Quick Reference): Guideline 1.1—Text Alternatives." *Web Accessibility Initiative*. World Wide Web Consortium. 29 Jan. 2019. https://www.w3.org/WAI/WCAG21/quickref/?versions=2.0&showtechniques=111%2C123%2C125%2C129#text-alternatives

Google Cloud. "Human labeling." *AI & Machine Learning Products documentation.* Google Cloud. 15 Nov. 2018. https://cloud.google.com/vision/automl/docs/human-labeling

Google Cloud. "Making an online prediction." *AI & Machine Learning Products documentation*. Google Cloud. 17 Apr. 2019. https://cloud.google.com/vision/automl/docs/predict

Keltz, Doug. "Understanding & Troubleshooting Closed Captions" (presentation). July 2014. https://www.smpte.org/sites/default/files/section-files/2014_July_Closed_Captioning.pptx

LeClere, Ellen. "As Libraries and Archives Digitize, Implications for Maintaining Individual Privacy." *MediaShift*. May 2016. http://mediashift.org/2016/05/as-libraries-and-archives-digitize-implications-for-maintaining-individual-privacy/

Markoff, John. "Seeking a Better Way to Find Web Images." *The New York Times.* 19 Nov. 2012. https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html

"MediaInfo." *MediaArea.net.* 2019. https://mediaarea.net/en/MediaInfo

"Micro-services." *Archivematica Development Wiki.* 14 Aug. 2015. https://wiki.archivematica.org/Micro-services

Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." National Institute of Standards and Technology, Special Publication 800-145. (2011).

MSV, Janakiram. "An Executive's Guide To Understanding Cloud-based Machine Learning Services." *Forbes*. 1 Jan. 2019. https://www.forbes.com/sites/janakirammsv/2019/01/01/an-executives-guide-to-understanding-cloud-based-machine-learning-services/#52d5709e3e3e

Otis, Jessica. @jotis13, Twitter post. 9 Mar. 2019. https://twitter.com/jotis13/status/1104410030458265600

Panetta, Kasey. "5 Trends Emerge in the Gartner Hype Cycle for Emerging Technologies, 2018." *Gartner, Inc.* 16 Aug. 2018. https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/

Schudel, Matt. "Henriette Avram, 'Mother of MARC,' Dies." *Information Bulletin*, Library of Congress. May 2006. Reprinted from *The Washington Post*, page B06, 28 Apr. 2006. https://www.loc.gov/loc/lcib/0605/avram.html

SMPTE 334M, as described by Sarkis Abrahamian in "EIA-608 and EIA-708 Closed Captioning." Evertz, n.d. https://evertz.com/resources/eia_608_708_cc.pdf

"Transcription time per audio hour: How long does transcribing really take?" *Opal Transcription Services.* Opal Transcription Services. Accessed 5 May 2019. https://www.opaltranscriptionservices.com/transcription-time-per-audio-hour/

"TRECVID data availability by year and task." *TRECVID*. National Institute of Standards and Technology. 17 Sept. 2018. https://trecvid.nist.gov/past.data.table.html

"Well-intentioned practice for putting digitized collections of unpublished materials online." *OCLC Research.* OCLC. Rev. 28 May 2010. https://www.oclc.org/content/dam/research/activities/rights/practice.pdf

"Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $77.6 Billion in 2022, According to New IDC Spending Guide." *IDC Corporate USA.* 19 Sept. 2018. https://www.idc.com/getdoc.jsp?containerId=prUS44291818