Chapter 10

More Than Just Algorithms:

A Machine Learning Club for Information Specialists

Mark Bell and Leontien Talboom

Introduction

Over the last few years, artificial intelligence (AI) and especially machine learning (ML) have become increasingly prominent in most industries, with the humanities sector being no exception. A number of institutions across the galleries, libraries, archives, and museums (GLAM) sector have been experimenting with and implementing algorithms or other computational techniques. Examples include The Living with Machines project at the British Library and the Alan Turing Institute, and The Museums + AI Network.[1] Specialised academic labs, like Yale's DHLab and Oxford's Visual Geometry Group, are also doing work related to ML.[2] At The National Archives (TNA) of the United Kingdom, activities include co-organising both the Computational Archival Science Network and an explainable AI workshop[3] as well as hosting an AI symposium for the archive sector and last year's Annual Digital Lecture on the topic of algorithmic bias.[4] TNA's Digital Strategy includes applying ML for appraisal, selection, and sensitivity review as well as improving access to the collections. It also emphasises the importance of developing "digital capability, skills and culture" within the organisation.[5]

Literature Review

A recent Archives, Access and AI conference showcased a number of projects that either use AI or aspire to do so.[6] The conference was not limited to speakers from the archives, as the title would suggest, but from organisations all across the humanities sector. For this chapter, the term *information specialist* will be used to refer to people from across this sector who acquire, appraise, and preserve materials. This group is starting to understand that their role with regard to computational methods, including ML, is changing. Public awareness of ML tools is also increasing. A 2016 survey by Ipsos Mori and the Royal Society found that only 9 percent of those surveyed recognised the term *machine learning*, yet 76 percent of respondents were aware of applications such as speech recognition and question answering, even if they didn't know they were powered by ML.[7] A similar poll in 2019 conducted by Kantar Public found that only 7 percent of respondents had never heard of AI, while 12 percent thought they knew a lot.[8] It should be noted that the term AI is far more ubiquitous in popular culture than ML, which may explain the disparity. Increasing awareness is partially explained by the increased automation in people's lives but also due to initiatives such as the Royal Society's Machine Learning project and the Finnish Government's Elements of AI course, which aims to educate 1 percent of European citizens on the basics of AI by 2021.[9]

There is also a growing societal awareness of what data can and cannot be used for, along with increased recognition of what can happen without the correct safeguards in place, such as amplifying existing biases within the data.[10] Mordell warns that all the social justice work done across the sector could be undone by the implementation of computational methods.[11] Griffey also warns about similar implications if these tools are not approached with caution.[12] Johnsson, Jakeway, et al. talk about how this technology is not only technical but also social and far more subjective than may be thought, as it relies on human judgment and biases.[13] It is not only about being cautious when implementing these tools; a number of papers have highlighted how critical the information specialist's skills—such as appraisal, selection, and cataloguing—are in the digital age.[14] Some researchers, however, argue that highlighting the problems and benefits is not enough and that there needs to be an emphasis on how important it is to engage in these discussions.[15]

Information specialists face other barriers when they implement these tools and experiment with them. Common barriers to AI projects include having insufficient data in the right format and insufficient skilled resources to take experimentation forward. As a result, growth can be witnessed in automated ML products, such as Google's AutoML,[16] which aim to democratise the building of models. The rationale is simple: data scientists are rare and therefore expensive and difficult to recruit, whereas there are millions of software developers already embedded into organisations.[17]

While processing a dataset of labelled example records (training data) through a proprietary "black box" algorithm is generally cheaper than hiring a data scientist, there is a loss of control over the process, including the ability to adjust the results and to explain the methods. Automated approaches take the focus away from the algorithms, which are hidden, and put it back firmly on the data that is used to train them. Information

specialists will become critical to the selection and creation of training data. This is a paradigm shift from a world where software developers elicit rules from users then design and develop a system that implements those rules. Subject matter experts now need to communicate with data scientists about selecting the right model that suits both the data and the application. These decisions are often a balance between accuracy and explainability.

Explainability is a rising trend in the debate around AI. Machine learning can be separated between statistical and algorithmic approaches, which Breiman describes as the "two cultures."[18] While both ultimately result in predictions, the statistical approach begins with identifying underlying models that describe physical phenomena, whereas algorithmic approaches are results-focused. Deep learning is used for complex tasks, such as image recognition and handwritten text recognition.[19] The "deep" part refers to the depth, or number of layers, in a neural network algorithm, each layer being a matrix of weights that are applied to the input data as it passes from layer to layer. The deeper the network, the more generalisable it becomes, but depth leads to greater complexity.[20]

Machine learning algorithms are evaluated against benchmark datasets, often termed the common task framework (CTF).[21] While this has led to incredible progress, many computer scientists are more focused on the performance of the tool than the understanding of how it functions,[22] which has led to some of the leading researchers in the field referring to it as "alchemy."[23] There have been attempts made to better understand neural networks with projects such as The AI Detectives.[24] Efron and Hastie consider empirical approaches like the CTF to be "ultimately unsatisfying without some form of principled justification."[25] While they are optimistic that the statistical inference community will eventually connect modern machine learning algorithms to a "central core of well-understood methodology," the issue remains that highly complicated tools are being built and the understanding of their internal workings is limited.[26] There is also the added problem that the benchmark datasets are not representative of the collections that information specialists would like to process with algorithms.

In order to address the growing interest in AI at TNA and the growing concerns around the ethical implications of these tools, a set of workshops entitled Machine Learning Club was organised. Members from all areas of the organisation attended the sessions. The aim of these workshops was not to turn information specialists into data scientists but rather to develop an understanding of what ML can offer to archives and which skills are needed to make the implementation of these tools successful. The sessions were designed to prepare staff to identify opportunities, remain alert to pitfalls, and be able to engage confidently with these exciting new technologies. This chapter explains how the Machine Learning Club was created and the content covered in the session. The knowledge and confidence gained from participating in the club are also discussed.

Machine Learning Club

The Machine Learning Club (MLC) was created to respond to a growing interest in the digital preservation team at TNA for more hands-on experience with ML. In order to pilot this initiative and measure the level of interest across the organisation, the authors

organised a series of lunchtime talks. These talks focused on different aspects of ML, starting with data preparation and a discussion of some well-known algorithms. Each talk lasted an hour and included further readings and examples of where to get practical experience. The club was well attended with around thirty attendees each session. Participants received homework assignments every session; however, there was little engagement with these assignments. Feedback received during the concluding session included a desire from participants to get more hands-on experience with this technology.

The authors initially selected Machine Learning Mastery and Towards Data Science as examples of online ML tutorials for the lunchtime talks.[27] These tutorials, however, can be highly technical and geared toward budding data scientists as well as require specific computer applications that present logistical challenges with regard to installation. The authors decided to modify the lunchtime talk sessions and create meaningful and relatable tutorials for information specialists. They presented these in four three-hour workshops offered monthly, providing aid and guidance when needed. Due to the large time commitment and the incremental nature of the workshops, the participants had to register and confirm with their manager that this time could be spent on MLC.

The organisers chose to use Google Colab as the environment to run the tutorials to avoid the technical issues surrounding installation of software applications and the lack of technical skills.[28] The Google-hosted "notebook" environments are run from the browser, making it unnecessary to install any software, and the computer code is run by clicking a play button in a cell.[29] The participants could therefore focus on the results given by the code and not on trying to get it to function. The tutorials and data were hosted on Google Drive, an environment familiar to most, which removed the barrier of trying something new. An example of one of the Google Colab tutorials from the MLC can be seen in figure 10.1.

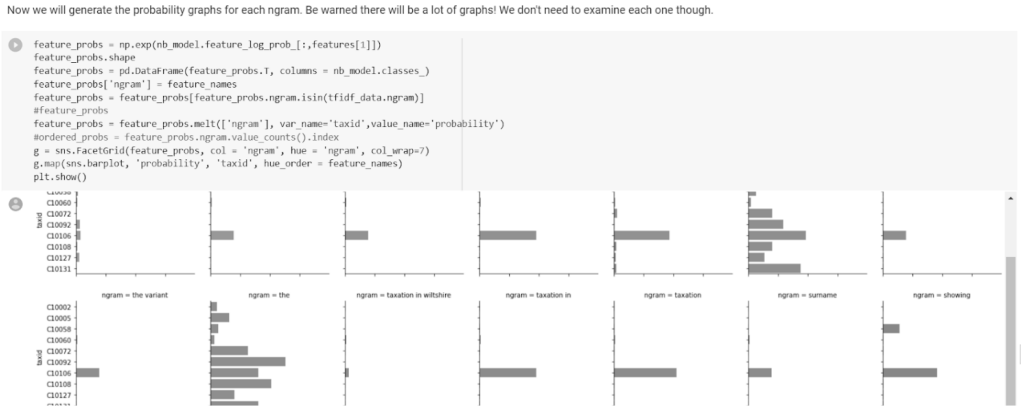

Figure 10.1

Part of an MLC tutorial, which is hosted on Google Colab. The first part of the tutorial consists of a piece of code that can be run by clicking the play button on the left-hand side. This will run the code and produce the underlying graphs as an output.

Session 1

The first MLC session focused on data analysis and data cleaning, which are important first steps in an ML workflow. The organisers started the session with a survey to measure the participants' knowledge of ML and to understand their motivations to take part in these workshops. They also used this information to modify the course materials for subsequent sessions. The presenters provided a general overview of the course and explained some basic ML concepts. The participants then took part in three tutorials, which involved different techniques and increasing levels of complexity.

The organisers selected a simple zoological dataset for the first tutorial.[30] Given that the dataset was in good condition, the presenters added a number of mistakes to determine if the participants could identify them. The tutorial focused on simple visualisation techniques, such as bar charts, for exploring numerical datasets. The second tutorial explored a US Census Income dataset, which was both numerical and categorical, making it more complex than the first dataset.[31] The participants learned statistical summarisation techniques, such as medians, averages, and quartiles, to aid them in understanding the data. This dataset also presented opportunities to discuss biases within the data. The final dataset, an Amazon food reviews dataset, was predominantly textual, which meant introducing word clouds and other text-mining techniques to be able to visualise this data, as the statistical techniques discussed in the other two tutorials would not be adequate for textual analysis.[32]

During this session, the presenters taught a basic understanding of the methods that can be used for different datasets and focused on the importance of data cleaning and gaining a thorough understanding of the data before using any type of ML. In order to encourage the participants to reflect on the activities and encourage wider group discussions, the code in the Google Colab notebooks was interspersed with questions.

Session 2

During the second session, the organisers provided the first hands-on experience of two ML algorithms: *nearest neighbours* and *decision trees*.[33] The participants began the workshop with a number of paper-based games (figure 10.2) to help build some intuition about how ML algorithms function. *Nearest neighbours* was introduced via a grid of playing cards, indexed by suit and number, and the participants had to decide whether to put down a blue- or red-backed card depending on the colour of nearby cards. To understand *decision trees*, the group was split into two and majority voting was used to build a tree for deciding whether to go outside depending on what the weather was. A third activity was used to explain the concept of *decision boundaries*,[34] where participants learned that an algorithm will change from predicting one class to another. Teams had to place pieces of strings to divide the squares and triangles plotted on paper; there was a penalty system in place (e.g., a shape being on the wrong side of the string) that replicated the optimisation process used by ML.

Figure 10.2
String, playing cards, and coffee: hands-on machine learning.

Following these exercises, there was a brief presentation to put them in context. The organisers then presented a number of tutorials to provide hands-on experience of the algorithms. In addition to the activities with the algorithms, the presenters explained two ways of assessing their performance: accuracy scores and confusion matrices. When participants needed more support, the presenters reviewed content from the previous session. The datasets used during this session were the same as the ones used in Session 1.

Sessions 3 and 4

Due to the COVID-19 pandemic lockdown, the organisers had to adapt the workshops to a new learning environment. They had originally planned to introduce an additional set of algorithms and compare them to the algorithms from Session 2. The organisers decided to keep this approach and include more content since they were no longer limited by the time restrictions of a classroom session. For this session, one data source was used: a set of categorised records from the Discovery catalogue at TNA.[35] The organisers chose this data source mainly because it is an internal dataset that is both familiar and relevant to most of the participants and because it is a prime candidate for future applications of ML. They created four tutorials that exemplified a small ML project, with each tutorial focusing on a specific aspect of the ML process: data analysis, data preparation, ML classification, and interpretation/explanation. For each tutorial, participants had to use data introduced in the previous section, meaning that data selection decisions made in the first part of the tutorial could influence the ML accuracy in the third part. The goal was

to demonstrate that choices made throughout the ML workflow could have an impact later in the process.

Session 3 is currently the last session of the ML Club, and the organisers hope to hold a fourth one once the situation around COVID-19 is resolved. They aim to focus the last session on the explainability of ML tools. In the meantime, the organisers have decided to postpone this session until it can be run in a classroom environment. They intend for the session to be heavily discussion-based, and their aim is to gather perspectives informed by the attendees' differing backgrounds. The intention is to break new ground in exploring the potential of ML from an information specialist viewpoint.

Discussion

Given the fact that these workshops were a new initiative, the organisers began Session 1 with a survey to gain a better understanding of the skills of the information specialists and their expectations with regard to the workshops. By combining the survey results with the participants' feedback during the sessions, the presenters were able to design each workshop around the needs of the attendees.

The organisers' decision to use Google Colab as the main environment was very beneficial. The environments could be run from any computer or device, without spending a lot of time downloading and installing software, which left more time for teaching and discussions. The information specialists were able to engage in the hands-on experience they were hoping for due to Google Colab. The attendees with no Python experience could still run code and interpret the results, while experienced coders could delve into the functions being used. Additionally, individuals with limited knowledge of Python attempted to experiment with the code, which was a positive and surprising outcome.

While the simple design and code initiation of Google Colab was a benefit, it was also its greatest drawback. Participants can simply press the activation button and not fully engage with the material. To solve this issue, the organisers included regular questions in the tutorials to prompt a more critical analysis of the visualisations or numerical summaries, and participants were also encouraged to work in groups and discuss amongst themselves. Group discussions had little success with most participants working in silence; however, participants were not afraid to ask questions. A participant in Session 1 asked, "What do you mean by algorithm?" which led to a discussion about the difference between standard algorithms being more analogous to recipes and the ML algorithms, which detect and learn from patterns in data.

To encourage group discussion and better exemplify ML algorithms, the organisers made changes to their approach for the second session. They created paper exercises and designed activities that required teamwork to test a wide range of parameter settings. Discussion was an important element of the classes and led to a number of interesting questions that showcased how the different perspectives of information specialists can be very beneficial to the ML conversation. They were particularly interested in data provenance, and their questions often highlighted the importance of contextualisation to information specialists, which would not necessarily be of concern to computer scientists.

The organisers noticed that the participants needed to understand everything, which led to several participants engaging in deep research about zoological classifications and causing some consternation when presented with an algorithm that was too complex to explain.

The exercises in the second session gave participants an opportunity to "think like an algorithm," and the presenters frequently referred to them in the sessions to enable participants to relate what was happening in Google Colab with their experience of placing a piece of string between points. Most participants had no issues with the level of Session 1, but Session 2 appeared to be more challenging for many. This discrepancy is likely due to the fact that the tools of data analysis (graphs, numerical summaries, word clouds) are familiar, whereas ML introduces a number of new and often abstract concepts. A number of participants have therefore taken the opportunity to revise the older material before commencing the Session 3 tutorials.

A participant in Session 3 expressed that they could not conceptualize the rationale for using an ML approach when there was no improvement of the results compared to the current manual process. Although the question can be addressed in terms of this specific archival problem, it highlights participants' expectations of AI technologies and shows that the techniques presented in a tutorial are not necessarily those used for real-world applications. At the same time as these workshops, two participants were involved in an AI project with external suppliers. They mentioned that it was useful to be able to employ their new knowledge in discussions with data scientists and to relate supplier presentations of their ML workflows back to the course materials.

After completing the sessions, a final survey was sent to participants to evaluate their final opinion on the MLC. The comments were overwhelmingly positive, with all respondents finding the MLC useful. Not all respondents were able to apply the material directly to their day-to-day tasks, but a number did see the benefit of being able to understand the basic concepts: "Even if I might not be directly applying the learning, it means that when I hear others talking about the topic, especially in professional environments, I have a much greater appreciation of what they are discussing and why it is significant."

Furthermore, respondents emphasised the importance of understanding ML for information specialists: "It's a necessity. As an archive, we will need machine learning to help us carry out our responsibilities." Another respondent agreed with this: "I can see so many applications for it in archives that I would consider it essential for anyone working (or planning to) work with digital records."

The organisers of the MLC received permission from TNA to create the workshops, and the time and space to hold the workshops were kindly provided. Unfortunately, not all organisations will have this flexibility; therefore, the Google Colabs have been made available.[36] The code may help others who are creating similar workshops or could act as an inspiration. As mentioned previously, few online tutorials may be as useful to information specialists; however, some examples aimed at the humanities sector could also act as an inspiration, such as the Programming Historian, the GLAM Workbench, the Archives Unleashed Project, and the CLARIAH Media Suite.[37] Each employs similar platforms to showcase the use of computational methods in the humanities.

Conclusion

The authors designed the Machine Learning Club to help information specialists understand what can and cannot be accomplished with ML; however, the goal was not to train future data scientists. Hopefully, participants have gained confidence and knowledge on this topic, which will make it possible for them to participate in discussions surrounding the implementation of ML across the GLAM sector. Moreover, participants may be more willing to join the conversation on archiving these methods and techniques for future re-use.

Information specialists have skills that are relevant to the explainable AI debate but often lack the confidence to participate in these conversations due to their lack of knowledge of the underlying computational methods. The authors are hopeful that the basic knowledge shared in the workshops will motivate the information specialists at TNA to stay engaged with this field and its opportunities while understanding its drawbacks. Machine learning has great potential within the GLAM sector and information management more generally, but it is also important to understand how it will impact information specialists. The club has hopefully made it clear that algorithms are only one part of the process and that people who understand records and data are as important as ever. While interesting and effective, perhaps AI is not the solution to everything.

Endnotes

1. "Living with Machines," The Alan Turing Institute, 2020, https://www.turing.ac.uk/research/research-projects/living-machines; Oonagh Murphy and Elena Villaespesa, "The Museums + AI Network," 2020, https://themuseumsai.network/about/.
2. "Digital Humanities Lab," Yale University Library, 2020, https://dhlab.yale.edu/; "Visual Geometry Group," University of Oxford, Visual Geometry Group, 2020, https://www.robots.ox.ac.uk/~vgg/.
3. Eirini Goudarouli, "Computational Archival Science (CAS): Exploring Data, Investigating Methodologies—Workshop Invitation," *The National Archives Blog* (blog), 2019, https://blog.nationalarchives.gov.uk/computational-archival-science-cas-exploring-data-investigating-methodologies-workshop-invitation/; Jenny Bunn et al., "Workshop on Human-Centred Explainable Artificial Intelligence," *UCL Blog* (blog), 2019, https://blogs.ucl.ac.uk/hexai/.
4. "Symposium 'Archives and AI' by FAN and the UK National Archives," International Council on Archives, 2019, https://www.ica.org/en/symposium-archives-and-ai-by-fan-and-the-uk-national-archives; Safiya Umoja Noble, "Annual Digital Lecture: Algorithms of Oppression" (Annual Digital Lecture, The National Archives, Kew, 2 April 2019), YouTube, https://www.youtube.com/watch?v=n3dQUYTN9PA.
5. The National Archives, *Digital Strategy 2017–2019* (London: The National Archives, 2017).
6. Lise Jaillant, "Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections," Poetry Survival, 2020, https://www.poetrysurvival.com/archives-access-and-ai-working-with-born-digital-and-digitised-archival-collections/.
7. The Royal Society, *Machine Learning: What Do the Public Think?* (The Royal Society, 2017).
8. Kantar Public, *AI PR Survey—Summary of Findings*, 2019.
9. The Royal Society, *Machine Learning*, 2020, https://royalsociety.org/topics-policy/projects/machine-learning/; EU2019FI, "Our Goal Is to Educate 1% of European Citizens in the Basics of AI," Elements of AI, 2020, https://www.elementsofai.com/eu2019fi.
10. Cathy O'Neill, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (London: Penguin, 2016); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018); Rana Foroohar, *Don't Be Evil: The Case Against Big Tech* (London: Penguin, 2019).

11. Devon Mordell, "Critical Questions for Archives as (Big) Data," *Archivaria* 87 (2019): 140–61.
12. Jason Griffey, "Conclusion," in "Artificial Intelligence and Machine Learning in Libraries, vol. 1," *Library Technology Reports* 55, 2019, 26–28.
13. Sylvester A. Johnson, "Technology Innovation and AI Ethics," in *RLI 299: Ethics of Artificial Intelligence, Research Library Issues* (Association of Research Libraries, 2019), 14; Eileen Jakeway et al., *Machine Learning + Libraries Summit Event Summary* (Library of Congress, 2020), 1.
14. Mary Lee Kennedy, "What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?," in *RLI 299: Ethics of Artificial Intelligence, Research Library Issues* (Association of Research Libraries, 2019), 3–13; William Kilbride, "Nothing About Us Without Us," *DPC Blog* (blog), January 20, 2020, https://dpconline.org/blog/nothing-about-us-without-us.
15. Chris Bourg, "What Happens to Libraries and Librarians When Machines Can Read All the Books?," *Feral Librarian* (blog), March 16, 2017, https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/; Catherine Nicole Coleman, "Artificial Intelligence and the Library of the Future, Revisited," *Stanford Libraries* (blog), November 3, 2017, https://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited; R. David Lankes, "Decoding AY and Libraries," *R. David Lankes* (blog), July 3, 2019, https://davidlankes.org/decoding-ai-and-libraries/; Annemaree Lloyd, "Chasing Frankenstein's Monster: Information Literacy in the Black Box Society," *Journal of Documentation* 75, no. 6 (2019): 1475–85; Michael Ridley, "Explainable Artificial Intelligence," in *RLI 299: Ethics of Artificial Intelligence, Research Library Issues* (Association of Research Libraries, 2019), 28–46.
16. "Cloud AutoML," Google, Google Cloud, 2020, https://cloud.google.com/automl.
17. Matthias Feurer, Aaron Klein, and Katharina Eggensperger, "Efficient and Robust Automated Machine Learning," *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2* (December 2015), 2962.
18. Leo Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science* 16, no. 3 (2001): 199–231.
19. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521, no. 7553 (2015): 436–44; Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* (2012), 1097–105; Guenter Muehlberger et al., "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study," *Journal of Documentation* 75, no. 5 (January 1, 2019): 954–76, https://doi.org/10.1108/JD-07-2018-0114.
20. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning Book* (Cambridge, MA: MIT Press, 2016), sec. 6.4.1, https://www.deeplearningbook.org/.
21. David Donobo, "50 Years of Data Science," 2015, 16.
22. Kiri L. Wagstaff, "Machine Learning That Matters," in *Twenty-Ninth International Conference on Machine Learning (ICML)* (Edinburgh, 2012), 529–36.
23. Ali Rahimi and Ben Recht, "Reflections on Random Kitchen Sinks," *Arg Min Blog* (blog), 2017, http://www.argmin.net/2017/12/05/kitchen-sinks/.
24. Paul Voosen, "The AI Detectives," *Science* (2017).
25. Bradley Efron and Trevor Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Cambridge, MA: Cambridge University Press, 2016), 447.
26. Efron and Hastie, *Computer Age Statistical Inference*, 447.
27. Jason Brownlee, Machine Learning Mastery, 2020, https://machinelearningmastery.com/; TDS Team, "Machine Learning," Towards Data Science, 2020, https://towardsdatascience.com/machine-learning/home.
28. "Welcome To Colaboratory," Google, Google Colab, 2020, https://colab.research.google.com/notebooks/intro.ipynb.
29. "Notebook Interface," Wikipedia, 2020, https://en.wikipedia.org/wiki/Notebook_interface.
30. Richard Forsyth, "Zoo Data Set," UCI Machine Learning Repository, 2017, https://archive.ics.uci.edu/ml/datasets/Zoo.
31. Ronny Kohavi and Barry Becker, "Census Income Data Set," UCI Machine Learning Repository, 1996, http://archive.ics.uci.edu/ml/datasets/Census+Income.
32. "Amazon Fine Food Reviews," Stanford Network Analysis Project, Kaggle, 2017, https://www.kaggle.com/snap/amazon-fine-food-reviews/.

33. "K-Nearest Neighbors Demo," Stanford Vision and Learning Lab, 2017, http://vision.stanford.edu/teaching/cs231n-demos/knn/; Jinda Shubham, "Decision Trees in Machine Learning," *Becoming Human* (blog), 2018, https://becominghuman.ai/decision-trees-in-machine-learning-f362b296594a.
34. Gaurav Kaushik, "Visualizing Decision Boundaries Builds Intuition About Algorithms," *Medium* (blog), 2018, https://medium.com/cascade-bio-blog/creating-visualizations-to-better-understand-your-data-and-models-part-2-28d5c46e956.
35. "Discovery," The National Archives, 2020, https://discovery.nationalarchives.gov.uk/.
36. Mark Bell, "Machine Learning Club," GitHub, 2020, https://github.com/nationalarchives/MachineLearningClub.
37. "About the Programming Historian," The Programming Historian, 2020, https://programminghistorian.org/en/about; Tim Sherratt, "Welcome to the Wonderful World of GLAM Data!," GLAM Workbench, 2020, https://glam-workbench.github.io/; Nick Ruest et al., "The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives," ArXiv:2001.05399, 2020; Liliana Melgar-Estrada et al., "The CLARIAH Media Suite: A Hybrid Approach to System Design in the Humanities," in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19* (Glasgow, Scotland UK: ACM, 2019), 373–77, https://doi.org/10.1145/3295750.3298918.

Bibliography

Alan Turing Institute, The. "Living with Machines." The Alan Turing Institute, 2020. https://www.turing.ac.uk/research/research-projects/living-machines.

Bell, Mark. "Machine Learning Club." GitHub, 2020. https://github.com/nationalarchives/MachineLearningClub.

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

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199–231.

Brownlee, Jason. "Machine Learning Mastery." Machine Learning Mastery, 2020. https://machinelearningmastery.com/.

Bunn, Jenny, Yvonne Rogers, Mark Bell, and Jo Pugh. "Workshop on Human-Centred Explainable Artificial Intelligence." *UCL Blog* (blog), 2019. https://blogs.ucl.ac.uk/hexai/.

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

Donobo, David. "50 Years of Data Science." 2015.

Efron, Bradley, and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge: Cambridge University Press, 2016.

EU2019FI. "Our Goal Is to Educate 1% of European Citizens in the Basics of AI." Elements of AI, 2020. https://www.elementsofai.com/eu2019fi.

Feurer, Matthias, Aaron Klein, and Katharina Eggensperger. "Efficient and Robust Automated Machine Learning." *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2* (2015), 2962–70.

Foroohar, Rana. *Don't Be Evil: The Case Against Big Tech*. London: Penguin, 2019.

Forsyth, Richard. "Zoo Data Set." UCI Machine Learning Repository. 2017. https://archive.ics.uci.edu/ml/datasets/Zoo.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning Book*. Cambridge: MIT Press, 2016. https://www.deeplearningbook.org/.

Google. "Cloud AutoML." Google Cloud, 2020. https://cloud.google.com/automl.

———. "Welcome To Colaboratory." Google Colab. 2020. https://colab.research.google.com/notebooks/intro.ipynb.

Goudarouli, Eirini. "Computational Archival Science (CAS): Exploring Data, Investigating Methodologies—Workshop Invitation." *The National Archives Blog* (blog), 2019. https://blog.nationalarchives.gov.uk/computational-archival-science-cas-exploring-data-investigating-methodologies-workshop-invitation/.

Griffey, Jason. "Conclusion." In "Artificial Intelligence and Machine Learning in Libraries," 1:26–28. *Library Technology Reports* 55, 2019.

International Council on Archives. "Symposium 'Archives and AI' by FAN and the UK National Archives." ICA, 2019. https://www.ica.org/en/symposium-archives-and-ai-by-fan-and-the-uk-national-archives.

Jaillant, Lise. "Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections." Poetry Survival, 2020. https://www.poetrysurvival.com/archives-access-and-ai-working-with-born-digital-and-digitised-archival-collections/.

Jakeway, Eileen, Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, Abigail Potter, and Kate Zwaard. *Machine Learning + Libraries Summit Event Summary*. Library of Congress, 2020.

Johnson, Sylvester A. "Technology Innovation and AI Ethics." In *RLI 299: Ethics of Artificial Intelligence*, 14–27. Research Library Issues. Association of Research Libraries, 2019.

Kantar Public. *AI PR Survey—Summary of Findings*, 2019.

Kaushik, Gaurav. "Visualizing Decision Boundaries Builds Intuition About Algorithms." *Medium* (blog), 2018. https://medium.com/cascade-bio-blog/creating-visualizations-to-better-understand-your-data-and-models-part-2-28d5c46e956.

Kennedy, Mary Lee. "What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?" In *RLI 299: Ethics of Artificial Intelligence*, 3–13. Research Library Issues. Association of Research Libraries, 2019.

Kilbride, William. "Nothing About Us Without Us." *DPC Blog* (blog), January 20, 2020. https://dpconline.org/blog/nothing-about-us-without-us.

Kohavi, Ronny, and Barry Becker. "Census Income Data Set." UCI Machine Learning Repository. 1996. http://archive.ics.uci.edu/ml/datasets/Census+Income.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* (2012), 1097–105.

Lankes, R. David. "Decoding AY and Libraries." *R. David Lankes* (blog), July 3, 2019. https://davidlankes.org/decoding-ai-and-libraries/.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature* 521, no. 7553 (2015): 436–44.

Lloyd, Annemaree. "Chasing Frankenstein's Monster: Information Literacy in the Black Box Society." *Journal of Documentation* 75, no. 6 (2019): 1475–85.

Melgar-Estrada, Liliana, Marijn Koolen, Kaspar Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martinez-Ortiz, Jaap Blom, and Roeland Ordelman. "The CLARIAH Media Suite: A Hybrid Approach to System Design in the Humanities." In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 373–77. CHIIR '19. Glasgow, Scotland UK: ACM, 2019. https://doi.org/10.1145/3295750.3298918.

Mordell, Devon. "Critical Questions for Archives as (Big) Data." *Archivaria* 87 (2019): 140–61.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, et al. "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study." *Journal of Documentation* 75, no. 5 (January 1, 2019): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Murphy, Oonagh, and Elena Villaespesa. "The Museums + AI Network." 2020. https://themuseumsai.network/about/.

National Archives, The. "Digital Strategy 2017–2019." London: The National Archives, 2017.

———. "Discovery." The National Archives. 2020. https://discovery.nationalarchives.gov.uk/.

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

———. "Annual Digital Lecture: Algorithms of Oppression." Presented at the Annual Digital Lecture, The National Archives, Kew, 2 April 2019. https://www.youtube.com/watch?v=n3dQUYTN9PA.

O'Neill, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin, 2016.

Programming Historian, The. "About the Programming Historian." The Programming Historian. 2020. https://programminghistorian.org/en/about.

Rahimi, Ali, and Ben Recht. "Reflections on Random Kitchen Sinks." *Arg Min Blog* (blog), 2017. http://www.argmin.net/2017/12/05/kitchen-sinks/.

Ridley, Michael. "Explainable Artificial Intelligence." In *RLI 299: Ethics of Artificial Intelligence*, 28–46. Research Library Issues. Association of Research Libraries, 2019.

Royal Society, The. "Machine Learning." The Royal Society. 2020. https://royalsociety.org/topics-policy/projects/machine-learning/.
———. 'Machine Learning: What Do the Public Think?' The Royal Society. 2017.
Ruest, Nick, Jimmy Lin, Ian Milligan, and Samantha Fritz. "The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives." ArXiv:2001.05399. 2020.
Sherratt, Tim. "Welcome to the Wonderful World of GLAM Data!" GLAM Workbench, 2020. https://glam-workbench.github.io/.
Shubham, Jinda. "Decision Trees in Machine Learning." *Becoming Human* (blog), 2018. https://becoming-human.ai/decision-trees-in-machine-learning-f362b296594a.
Stanford Network Analysis Project. "Amazon Fine Food Reviews." Kaggle, 2017. https://www.kaggle.com/snap/amazon-fine-food-reviews/.
Stanford Vision and Learning Lab. "K-Nearest Neighbors Demo." 2017. http://vision.stanford.edu/teaching/cs231n-demos/knn/.
TDS Team. "Machine Learning." Towards Data Science. 2020. https://towardsdatascience.com/machine-learning/home.
University of Oxford. "Visual Geometry Group." VGG—University of Oxford. 2020. https://www.robots.ox.ac.uk/~vgg/.
Voosen, Paul. "The AI Detectives." *Science* (2017).
Wagstaff, Kiri L. "Machine Learning That Matters." In *Twenty-Ninth International Conference on Machine Learning (ICML)*, 529–36. Edinburgh, 2012.
Wikipedia. "Notebook Interface." Wikipedia. 2020. https://en.wikipedia.org/wiki/Notebook_interface.
Yale University Library. "Digital Humanities Lab." DH Lab. 2020. https://dhlab.yale.edu/.