The goal is to assess the model's breadth and depth of knowledge across many different subjects and its ability to apply this knowledge effectively.

6.3.7 Coding

Executive summary: some LLMs are extremely good at writing computer code.

Several benchmarks are widely used for evaluating the coding capabilities of large language models. These benchmarks assess various aspects of programming skills, including code generation, bug fixing, and code completion. Some of the most notable benchmarks include:

1. **CodeXGLUE**:
    - **Description**: A comprehensive benchmark and collection of datasets for code-related tasks, including code generation, code completion, code summarization, and more.
    - **Tasks**: Code-to-code tasks, text-to-code tasks, code-to-text tasks, etc.
2. **HumanEval**:
    - **Description**: A benchmark specifically designed to evaluate the functional correctness of generated code by using unit tests.
    - **Tasks**: Given a natural language prompt, the model must generate a correct and functional code snippet that passes the provided unit tests.
3. **APPS (Automated Programming Progress Standard)**:

- **Description**: A benchmark that includes a diverse set of coding problems, ranging from introductory to complex algorithmic challenges.
- **Tasks**: The model must write code to solve given programming problems, and solutions are evaluated based on correctness and efficiency.

4. **MBPP (Mostly Basic Programming Problems)**:
   - **Description**: A dataset of basic programming problems designed to evaluate the model's ability to generate correct and executable code.
   - **Tasks**: Code generation based on problem statements, with evaluations focusing on correctness and simplicity.
5. **CodeBERT**:
   - **Description**: A pre-trained model for programming languages, evaluated on a variety of coding tasks.
   - **Tasks**: Includes tasks like code search, code documentation generation, and code completion.
6. **XLCoST (eXtreme Language Code Search and Translation)**:
   - **Description**: A benchmark for code search and code translation across multiple programming languages.
   - **Tasks**: Code search, where the model retrieves relevant code snippets based on natural language queries, and code translation, where the model translates code from one programming language to another.

These benchmarks provide a robust framework for evaluating the performance of language models in coding tasks, helping researchers and

developers understand the strengths and weaknesses of their models in real-world programming scenarios. [This section was written by an LLM.]

6.4 Artificial General Intelligence (AGI)

There is the idea of Artificial General Intelligence, which roughly amounts to simultaneously being more intelligent than a human at all intelligent endeavors. Quite what this notion is exactly has never been tightly pinned down. Supposedly OpenAI have an internal 5-level scale to track progress towards AGI. It is:

1. Conversation to the standard of present day chatbots
2. Solving problems to the level of a person with a PhD.
3. Being capable of taking actions on a user's behalf.
4. Creating new innovations.
5. Performing the work of entire organizations of people.

Present LLMs can do 1, and nearly do 2, but are yet to reach the other three levels (Metz 2024). There have been other suggestions on AGI. We will follow one of them, that from François Chollet (Chollet 2019; ARCPrize 2024).

Chollet invites us to consider intelligent tasks, as examples playing chess, summarizing a document, and solving a High School math problem. We know from current LLMs and their benchmarks that very likely AI systems will be able to do all such tasks far better, far more 'intelligently', than a human. Would this mean that AGI had been created? Chollet answers No.

Chollet observes that designers of a program to play chess know exactly what the problem is. Similarly, creators of AI systems to summarize documents know what is to be done. He suggests:

> Measuring task-specific skill is not a good proxy for intelligence.... Intelligence lies in broad or general-purpose abilities; it is marked by *skill-acquisition* and generalization, rather than skill itself.
> **AGI is a system that can efficiently acquire *new* skills outside of its training data....**
> This means that a system is able to adapt to a new environment that it has not seen before and that its creators (developers) did not anticipate. (ARCPrize 2024).

Chollet suggests that the way LLMs work is that they are large 'interpolative memories'. They have seen, and remembered, a vast quantity of facts, data, and patterns— all of the Internet, basically. Then they retrieve or fill in the gaps, to produce their answers from prompts. But, Chollet observes, human intelligence is of a more general kind. The world is always changing. Humankind has need not only to deal with the familiar, but also to confront the totally novel. Even a five-year-old can solve problems both that they have never seen before and of a *kind* that they have never seen before. Seemingly LLMs cannot do this.

Chollet has designed an 'IQ test for Artificial General Intelligence'. Five year olds can do it to a level of 50%, unexceptional adults can score about 85% on it, and, mid-2024, the best LLMs might get around 10% (Patel 2024).

6.5 The ARC-AGI Benchmark

Chollet asserts that ARC-AGI is the only current AI benchmark that measures progress towards general intelligence.

The benchmark is a whole suite of tests. They are designed to be resistant to prior memorization. Earlier tests won't help you. Nor will study of any matters whatsoever. Here is a sample. You are given three examples of inputs and their corresponding outputs. Then there is a test input for which you must suggest the output. Try it. (No prizes for success— you should be able to do it. Many five-year-olds can.)

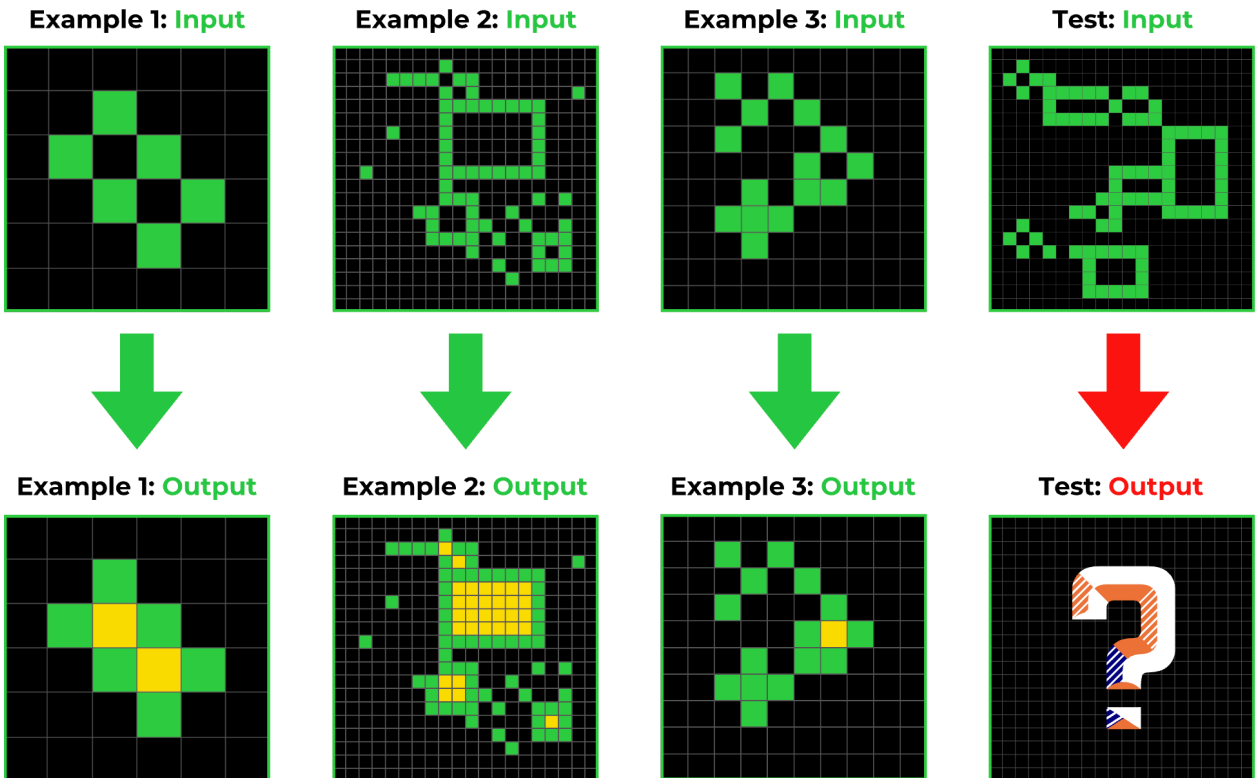

Figure 17. Sample ARC-AGI test

As mentioned, humans would score about 80% on the entire benchmark. LLMs might get 10% or so. Chollet and Mike Knoop have offered a substantial monetary prize to any AI system that can score 85% i.e. that is be better than humans (Patel 2024).

Mid-2024, it looks as though the ARC challenge poses a real difficulty for LLMs. It is, of course, possible for LLMs to improve. It is also possible for some AI or machine learning system which uses different principles or techniques from LLMs to succeed with the ARC test. What the test is looking for is a system that can adapt to a task that is truly novel from the perspective of its training data (Patel 2024).

6.6 Artificial Super Intelligence (ASI)

There is the notion of the 'singularity', which was introduced by Ray Kurtzweil (Kurzweil 2005). The idea here is that as technology advances there will come a point when it becomes smarter than we are. At that point, the 'singularity', the machines can simply design themselves and become smarter and smarter in a runaway fashion. There is the similar idea in AI and that is Artificial Super Intelligence (ASI). If indeed Artificial General Intelligence (AGI) is possible, then, presumably these intelligent systems can simply design even better machines— better systems than humans can design, and better systems than they themselves are. The result would be Artificial Super Intelligence (ASI). More than a few 'futurists' are very concerned about the possibility of ASI. An ASI system would have no need for humans, and hence might be a threat to humans. Further, any

individual, group, or country that was the sole possessor of ASI would, or might, have unlimited power and control over everyone else.

A detailed argument to that effect is provided by Leopold Aschenbrenner in his 160 page book *Situational Awareness: The Decade Ahead* (Aschenbrenner 2024). This might be summarized as follows (with brief annotations):

- There will be AGI by about 2027. [Anschenbrenner gets this by extrapolation from current rates of progress. But, for example, if System 2 thinking proves a barrier, i.e. the ARC-AGI consideration, this reasoning might not be sound.]
- There will be ASI about a year later. [This comes from the suggestion to create thousands of AGI bots and throwing them at every unsolved scientific, mathematical, and other problem. But, so-to-speak, having a thousand 'Einsteins' might not get us any further than having just one of them (cf. *The Mythical Man-Month* (Brooks 1975).]
- ASI might convey a decisive military and political advantage to whoever has it first. The 'might' here mainly concerns alignment, or 'superalignment'. ASI is of advantage to its owner or creators only if ASI instances do what the owners want them to. If ASI instances 'have minds of their own' and do whatever they wish, that may make them useless, or even dangerous, to their owners. [But this alignment problem is highly non-trivial because the systems are so large and complex that humans, or RLHF, simply cannot understand what is going on. Reinforcement learning with human feedback (RLHF) just will not work. There will need to be another way.]

- What characterizes these deep learning, LLM, or even ASI, systems are the weights they use in their models. Weights are just numbers. There may be many of them, billions, or trillions, but they are still just numbers. If an adversary can steal the numbers, they can create the systems without bothering to do the research. [True, with evidence.]
- Bad actors at the state level (say, Russia, China, Iran, or North Korea) likely could steal the weights from any ordinary commercial enterprise without much difficulty. Thus, there is a serious security problem. [Probably True, and there is evidence (Stuxnet's sabotage of Iran's nuclear centrifuges and, separately, Pegasus spyware against smartphones.)]

6.7 Annotated Readings for Chapter 6

Aschenbrenner, Leopold. "Situational Awareness: The Decade Ahead," 2024. https://situational-awareness.ai/.

Patel, Dwarkesh. "Leopold Aschenbrenner - China/US Super Intelligence Race, 2027 AGI, & The Return of History," 2024. https://substack.com/home/post/p-145136502.

Patel, Dwarkesh. "Francois Chollet, Mike Knoop - LLMs Won't Lead to AGI - $1,000,000 Prize to Find True Solution," 2024. https://www.dwarkeshpatel.com/p/francois-chollet. (Patel 2024). This is a 90minute video podcast featuring a discussion between Chollet and Patel. There is a written transcript. (Dwarkesh Patel's podcasts are excellent.)

Chapter 7: Bias and Unfairness

7.1 Algorithmic Pipeline + Data = Machine Learning

Niklaus Wirth's 1976 book *Algorithms + Data Structures = Programs* is one the most important and influential books in computer science. It led to the style of structured programming, the development of the Pascal programming language, the move toward typed programming languages, and the design of many University programming courses.

Somehow, nowadays, the whiff of the title has found its way into modern characterizations of ML. Many say that *Algorithms + Data = Machine Learning*. Then reasoning proceeds from the premise 'There is (plenty of) bias in Machine Learning' to 'There is bias in ML Algorithms and there is bias in ML Data.' This is not quite right, though. It is not right on the location of bias (and locating the bias correctly will help us to address it). When a computer program is written there is the question of what the program is supposed to do. Is it supposed to add up some numbers? Is it supposed to find the address of someone in a Contacts book? Is it supposed to suggest folk qualified for a mortgage on the basis of demographic information about them? This what-it-is-supposed-to-do part is usually called the *specification*. What a specification amounts to varies with circumstance. A hobbyist programmer may have a rough mental idea of what she is trying— that may be a specification without anything being written down. In contrast, a specification for a program, or project, like Google Documents may consist of hundreds of pages of text written in a

very formal style. Specifications can change as projects develop and are in process (for example, to omit features that prove to be difficult to implement). But changing specifications is considered bad form, and it is usually avoided if possible.

Imagine this as an example of some biased software. A mortgage company gives their expert programming team the task of producing some mortgage qualifying software with a partial specification that only black applicants should qualify. The programmers, expert as they are, then produce a flawless program to do exactly this. The outcome may be biased. Let us suppose that it is. Where does the bias come from? It may come from the data. But suppose that the data, its veridicality, its sampling, etc., is perfect in every way. So, the bias has not come from the data. What about the algorithms in the program? Well, it could easily be that they are entirely perfect in every way. So, there is no bias there either. What is left? The bias comes from the specification.

A formal specification is only part of the programming infrastructure surrounding projects (especially so in large organizations, businesses, or institutions). ML projects are mostly complex. There often is development and deployment. There is an entire pipeline, a programming 'environment'. Bias can arise anywhere in this, or, indeed in several different places. Johanna Johansen et al. suggest the label 'programming artifacts' for this infrastructure (Johansen, Pedersen, and Johansen 2021). This is a good idea. However, many ML researchers and programmers tend to use a flowing water metaphor to capture the process. They talk of 'downstream',

'upstream', and 'pipeline', and other similar descriptive nouns. We will do the same.

Algorithmic Pipeline + Data = Machine Learning.

Bias in Machine Learning comes from bias in the Algorithmic Pipeline or bias in the Data.

Some commentators allow the location of bias to spread beyond the individual algorithmic pipelines to the AI industry as a whole (for example, that there is a preponderance of male employees, that much of it is funded by the state and the military, that it is commercial aiming to make a profit). (See, for example, (de Hond, van Buchem, and Hernandez-Boussard 2022).)

7.2 Some Clarification of the Terms 'Bias' and ‘Unfairness’

There is a need for care when using the terms ‘bias’ and ‘unfairness’ in the context of machine learning. Most educated adults know what these words mean in the sense of being able to produce illustrative sentences that use these words correctly and being able to paraphrase the sentences of others that use the words. In machine learning, some of the literature uses these two words interchangeable as synonyms (Pagano et al. 2023). This is not correct in general, though. The word ‘bias’, or the phrase ‘predictive bias’, have extensive uses in statistics and machine learning to mean ‘error’ or ‘systematic error’. But many of these errors are not either fair or unfair.

Imagine an ML program to predict the weather and suppose the weather could be only either sunny or rainy. Suppose the model's daily predictions were sometimes correct and sometimes mistaken. (That might be the best one could hope for.) But if the model predicted 100 days of rain in the year and actually there were 300 observed days of rain that year, the model has predictive bias. This kind of bias has nothing to do with unfairness to anything or anybody. It is not unfair to rainy days. ML researchers would like rid of this kind of predictive bias from their model. But this is not an ethical mandate. It is not a matter of justice. The researchers just want their models to be more accurate. Here is a second example. LLMs predict the next letter, token, or word from a context or prompt. Imagine that GPT-0.01, working in English, never predicted the letter 'e' as being the next letter. GPT-0.01 would have predictive bias. But its predictions are not unfair. (Although, the children's television program Sesame Street might say they are unfair to the letter 'e'.) There are also harms that can result from the predictions of machine learning programs. But there can be harms without bias (where there are no errors in the program and its predictions) and harms without unfairness (where every person or group is harmed equally).

Tiago Pagano and fellow authors write:

> Prediction-based decision algorithms are being widely adopted by governments and organizations, and are already commonly used in lending, contracting, and online advertising, as well as in criminal pre-trial proceedings, immigration detention, and public health, among other areas.
> However, as these techniques gained popularity, concerns arose about the bias embedded in the models and how fair they are in

> defining their performance for issues related to sensitive social aspects such as race, gender, class, and so on.
> Systems that have an impact on people's lives raise ethical concerns about making fair and unbiased judgments. As a result, challenges to bias and unfairness have been thoroughly studied, taking into consideration the constraints imposed by corporate practices, legislation, societal traditions, and ethical commitments. Recognizing and reducing bias and unfairness are tough undertakings because unfairness differs between cultures. As a consequence, the unfairness criteria are influenced by user experience, cultural, social, historical, political, legal, and ethical factors (Pagano et al. 2023).

Bias is a huge archipelago of topics. The word 'bias' has several totally different meanings.

In the ML technical core, there is bias in the context of the weighting of inputs to software neurons in neural nets. Then, in wider ML, Aylin Caliskan et al. define bias with the following statement:

> *In AI and machine learning, bias refers generally to prior information, a necessary prerequisite for intelligent action.* Yet bias can be problematic where such information is derived from aspects of human culture known to lead to harmful behavior. Here, we will call such biases "stereotyped" and actions taken on their basis "prejudiced." (Caliskan, Bryson, and Narayanan 2017) [Italics added.]

This is important. It is completely standard in the context of machine learning. However, it is completely non-standard, and totally at odds with what ordinary people might mean by bias in ordinary settings. Generally, bias is not a good thing, and we would like rid of it. It is to be spurned. But, 'information' can mean 'knowledge' (Frické 1997). So, in a ML program, any

part of the entirety of human knowledge might be 'bias'. For example, that 2 + 2 = 4 might be bias.

Machine learning researchers also often use the term ‘bias’ in connection with the predictions of a model. More specifically, they might use the phrase ‘predictive bias’ in this setting. Many ML models make predictions. Then, of course, the question arises of whether the predictions are correct or incorrect. If they are incorrect, especially incorrect in a systematic way, the model would be said to exhibit ‘predictive bias’.

Separately, also related to ML— with causal diagrams, and the statistics of causality, there is the central notion of confounds. These are often called 'bias'.

In chance like set-ups, typically for gambling, such as roulette wheels, rolled dice, or tossed coins, the set-up is unfair or biased if the chances are not as they should be. If a thrown coin favors Heads over Tails, it is biased.

In the context of people and diversity, there is the notion of bias meaning 'unfair prejudice'. For example, The Office of Diversity and Outreach of the University of California San Francisco offers this description of bias in a general non-computing setting:

> **Bias** is a prejudice in favor of or against one thing, person, or group compared with another usually in a way that's considered to be unfair. Biases may be held by an individual, group, or institution and can have negative or positive consequences. There are types of biases 1. **Conscious bias** (...**explicit** bias) and 2. **Unconscious** bias (... **implicit** bias)

> … biases, conscious or unconscious, are not limited to ethnicity and race. … biases may exist toward any social group. One's age, gender, gender identity physical abilities, religion, sexual orientation, weight, and many other characteristics are subject to bias. (UCSF Office of Diversity and Outreach UCSF 2022)

[Of value as background here on this sense of bias are Project Implicit, the Implicit Association Test (IAT) and the work of the Kirwan Institute ("Project Implicit" 2011; Kirwan Institute 2017).]

Then there are cognitive biases. One example is confirmation bias. This is the tendency, with beliefs or knowledge, for people to seek out, or give more weight to, evidence or arguments that support or 'confirm' views or opinions that they already hold (Wikipedia 2023b). Another example of a cognitive bias is that actual human reasoning, both the principles used and the individual instances of it, is often incorrect, maybe even almost always incorrect (A. Tversky 1974; Kahneman 2011). One famous instance of this is the base-rate fallacy embodied in the so-called Harvard Medical School test ((Casscells, Schoenberger, and Graboys 1978) see also Appendix C).

Further, the conceptual schemes and natural languages that are in use reflect all sorts of attitudes, and attributions of accidental features that do not really belong in an accurate description of what they are applied to. There is bias in conceptual schemes and language. Unfortunately, more than a little ML, especially unsupervised, or self-supervised, learning (i.e. finding patterns and clusters where there are no sample right answers), builds off the Natural Language Processing (NLP) of books, recordings, language, and conceptual schemes. NLP needs a section to itself (which we will get to).

7.3 Forms of Bias in Wider Machine Learning

Kate Crawford, in her 2017 keynote address to the Neural Information Processing Systems Conference, identifies three main forms of bias in the context of ML: harms of allocation, harms of representation, and harms of classification (Crawford 2017; Barocas et al. 2017). The first concerns who does or does not get the mortgages, or who does or does not get shorter prison sentences when re-offending, etc. i.e. fairness of allocation. The second concerns how individuals, groups, or even things and classes of things, are represented or portrayed or named. [This involves emotive content, which is a topic introduced in Appendix A.] Of course, being represented in a negative way may have consequences, for example, that of not being allocated a mortgage. The third concerns how humans, individual human beings or groups of human beings, are watched, perhaps surveilled, and classified, usually for other, often discriminatory, purposes, for example, for apartheid as it was in South Africa (Bowker and Star 2000; Gandy Jr. 2021; Crawford 2022). (An allusion here is to the *Panopticon* of Jeremy Bentham and his brother, Samuel Bentham, (cf. Foucault's panoptic prison (Brunon-Ernst 2012))). We should note that this kind of classification is different to the classification done by librarians. Librarians classify recorded documents and information resources (e.g. books), not human beings.

As noted, bias is a vast territory. Even within ML it is possible to expand Crawford's classification of three biases out to seven or more biases (see, for

example, (Suresh and Guttag 2021)). Also of note, Su Lin Blodgett et al. critically surveyed 146 papers on bias in NLP and found that:

> ...the majority of them fail to engage critically with what constitutes "bias" in the first place (Blodgett et al. 2020)

The Blodgett et al. paper does have valuable suggestions. In part, first, that harms of allocation, and harms of representation will take you a long way when considering bias. Then:

> ... work analyzing "bias" in NLP systems should provide explicit statements of why the system behaviors that are described as "bias" are harmful, in what ways, and to whom [further text omitted here]. (Blodgett et al. 2020)

[Bommasani et al. is another important source on the topic of bias in ML (Bommasani et al. 2022)]

For our purposes, and as a practicality, we can restrict ourselves going forward primarily to fairness, representation, and classification (primarily in the librarian's sense of 'classification').

7.4 Bias in Natural Language Processing

Recently, say since about 2017-2018, NLP has become a huge and significant part of ML. This is because of the emergence of Large Language Models and Foundation Models (which are being discussed in more detail elsewhere). These models form the core of many of the truly innovative

modern systems (hence 'Foundation Models'). In turn, they are based on natural language processing (NLP). So, NLP has become more important than ever, and biases in NLP can leak into the modern innovations.

A well-known and introductory example of apparent bias in NLP concerns the translation of Turkish. Turkish does not have gender pronouns, so translating into Turkish can lose the gender of the original. Then translating back may use 'gender bias' to make a guess as to the gender of the pronoun. A few years ago, you used to be able to do this on Google Translate:

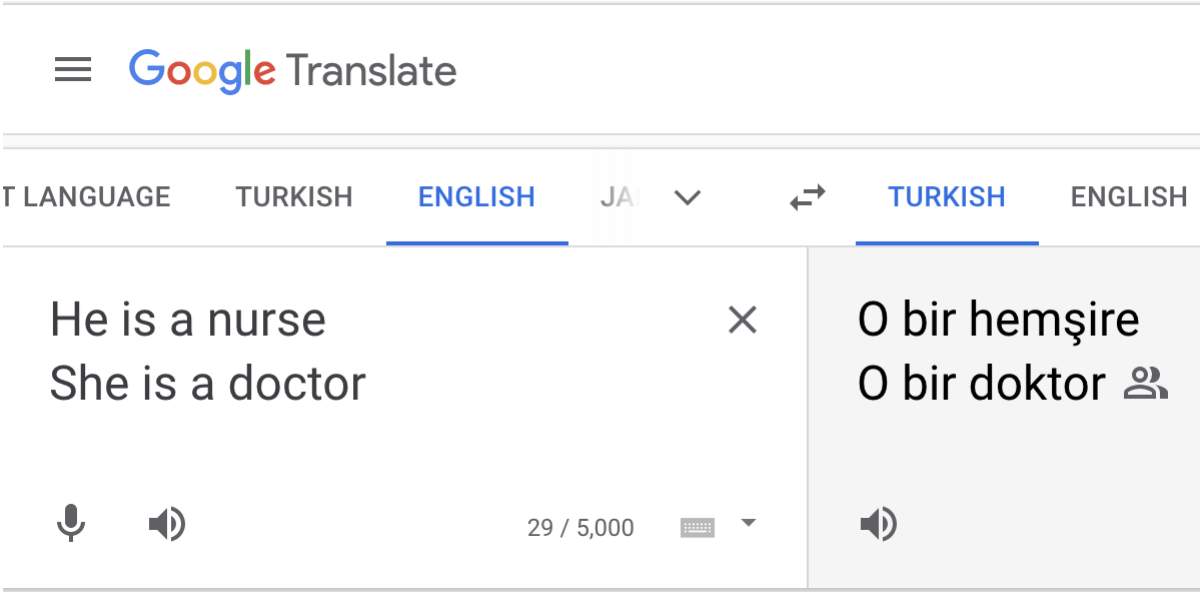

Figure 18. Translating from English to Turkish.

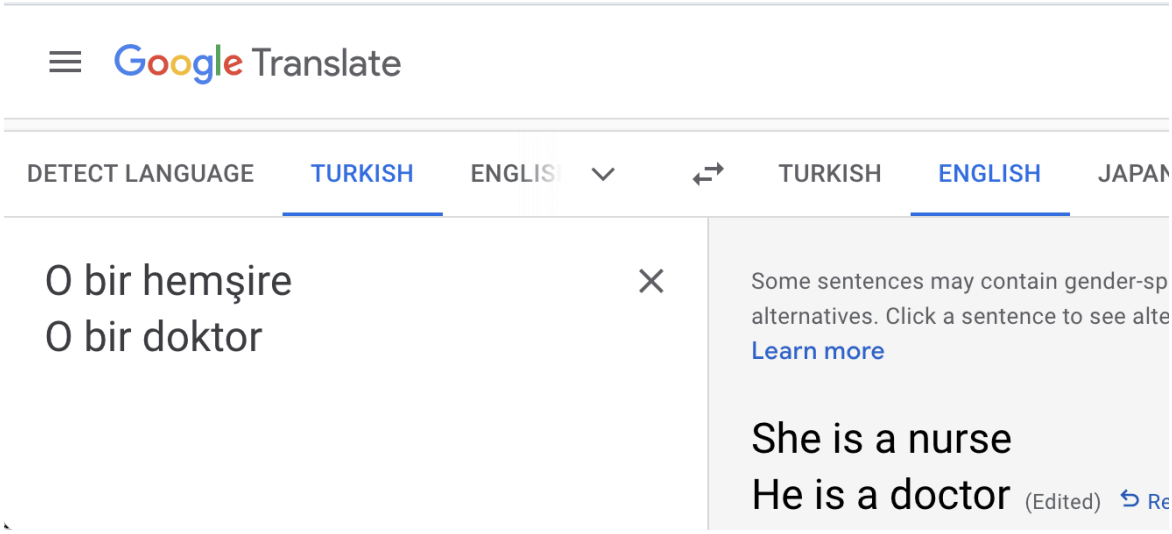

Figure 19. Translating the translation back from Turkish to English.

There are a few points to be made. A human translator would make the same 'mistake'. There is no context in the brief Turkish text to pick up the gender of the doctor and nurse in question. Given a longer text, say a magazine article or a novel, both the human translator and the ML translator would get this right. Separately, nowadays, as you can see from the screen shot, Google translate alerts the User to the gender-specific alternatives. Finally, it is not entirely clear that this kind of example is a case of bias. There are more male doctors than female doctors, presumably more male Turkish doctors than female Turkish doctors. There is a higher probability of a doctor being male than being female. We may find that fact unfortunate, and not good for society, for women, for medicine, and for the good life in general. But it is a fact. Consequently, if presented with those probabilities and a remote doctor, of unknown male or female gender, unbiased reasoning would suggest the conjecture that the doctor was male.

(In the absence of other information, you should choose the base-rate as your probability. [There is more on the base-rate in Appendix C.])

Research, for example that of Aylin Caliskan et.al., has shown that everyday languages have biases built in, and those biases can seep into the results of ML (Caliskan, Bryson, and Narayanan 2017; Caliskan 2021). Aylin Caliskan et.al. write

> Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology (Caliskan, Bryson, and Narayanan 2017).

The biases already exist in ordinary languages. ML did not create these biases; it just identifies them. Notice here the distinction they make in the second and third categories between the problematic and the veridical. This can be illustrated with the Turkish doctor case. It seems that, as a matter of fact, ordinary everyday English, in English societies and cultures, has the status quo bias that doctors are male (that is why the example translation from Turkish goes wrong). Separate from this is the question of whether this bias is problematic— whether we *should* assume that doctors are male— and probably most people would say that we should *not* assume this. Now, there is here a gulf between facts and values, between what biases *do* exist (the veridical) and what biases *should not* exist (the problematic). Once values enter there are further problems. What is the reasoning, the evidence, and the motivations for decisions on values? Who

decides? And on what basis? Then, if a view can be formed, how could it be implemented, either in natural language or in ML software? Some moral positions can have the backing of the law— murder is both wrong and illegal. But we presumably would not want to invoke the law against the bias that doctors are male. In brief, there are many problems. To continue. These biases are in natural languages, and we are immersed in these languages. Likely there is some two-way traffic between the languages we use and the biases we have. Our linguistic practices do change over time— we, in English speaking America, are no longer comfortable with phrases like 'yellow peril' or words like 'ni**er'. That our linguistic biases have changed does not mean that shortly we will be free of all linguistic bias. Some fear that ML will amplify or entrench the existing biases in natural languages. It is hard to know. One factor that is awkward here is that much of NLP ML is unsupervised or self-supervised. That means that the programs are often not being told the 'right' answers. The language corpora that they work on are almost always huge. The GPT-x series, for example, essentially scan the entire Internet, maybe trillions of word tokens. If a program is looking for patterns, undirected, through the whole of the Internet, it is hard to see how it could omit biases from that search (it does not even 'know' that they are biases). (GPT-3 itself, for example, can be given prompts, which can give it some direction. Prompts can tell it to be safe and not to be toxic or biased.) Some biases can be reduced or removed in ML software on a piecemeal basis. There are de-biasing initiatives (e.g. (Bolukbasi et al. 2016)). We should take into account here considerations of offensive speech, hate speech, and legal protections of free speech (such as the First Amendment in the United States). Google, and similar large outlets of speech, have policies and guidelines on being parties to the

publication of hate or offensive speech. Basically, the policies respect the laws while keeping themselves clear of what might be marginal cases. The biases that occur everywhere in everyday language would not be front and center. Some tentative conclusions are… ML needs to use NLP to produce translations, sound interfaces, verbal assistants, and so forth. These technological possibilities are, on balance, so valuable that it is hard to imagine not pursuing them. Then data from NLP will likely contain bias, and that bias will be hard to address.

7.5 Some Clarification of the Term 'Algorithm'

Presumably, some ML programs are biased (just being open minded here on what the word 'bias' might mean in this context). But we need to be measured in addressing a serious issue. Here is what some prominent commentators write:

> Algorithms are neither neutral nor objective. They are programmed by human beings, who have both conscious and unconscious biases, and those biases are encoded into software (Cordell 2020).

> … essentially a lie— namely, that algorithms were being presented and marketed as objective fact. A much more accurate description of an algorithm is that it's an opinion embedded in math (O'Neil 2018).

> … algorithms are the result of human endeavor and human-generated data sets so they are just as biased as we are. We just can't see it.
> As humans, we all have implicit biases. And as we build these new systems – facial recognition, AI, analytical algorithms – we're

> creating them in our own image, with these biases baked in. (Ayre and Craner 2018).
>
> … algorithms are the product of explicit and latent biases held by humans (Padilla 2019).

The sentiments expressed here are both factually wrong and pernicious. The Ryan Cordell passage, to take one example, is from a report for the Library of Congress, which is the most important institutional body in American librarianship. The Library of Congress here thus presumably approves of, and certainly promulgates, a report that misleads librarians.

Most algorithms, computer science algorithms and folklore 'algorithms', are not biased. [See Section 1.7. For example, the algorithm division by successive subtraction is not biased, period.] Separately, the argument 'All humans are biased, therefore, all human products (e.g. software) are biased' is invalid and has a false premise. Some more detail, or evidence, can be added here. The US federal agency IMLS (Institute of Museum and Library Services) has funded a useful and informative educational resource on algorithmic awareness, aimed to an audience of information professionals (Clark [2018] 2022). The resource mentions, and demonstrates, the following as important algorithms, used in online search:

> … PageRank, merge sort and heap sort, Dijkstra's algorithm, link analysis, and TF-IDF (Term Frequency-Inverse Document Frequency) (Clark [2018] 2022).

None of these algorithms is biased. Take Dijkstra's algorithm, for example. As an analog of what it does: it will find the shortest path, or route, between

any two cities, where there are several cities connected by roads (sometimes a direct route is shortest, sometimes going via intermediary cities is the shortest). This algorithm is not biased (in any sense of 'bias' whatsoever). Here is a general argument to refute the view that all algorithms are biased: Dijkstra's algorithm is an algorithm, Dijkstra's algorithm is not biased, *ergo*, not all algorithms are biased.

We do not wish to get tied up here with arguing the meaning of words. If those concerned with shortcomings with the use of computers in society regularly talk about 'algorithmic bias', the 'social power of algorithms', or even '#FuckTheAlgorithm', that is fine (Beer 2017; Benjamin 2022). We will open our minds to this. For ourselves, we prefer 'bias in the programming artifacts', or 'bias in the algorithmic pipeline' or just the plain 'bias in the software'. We will cautiously and tentatively use this, and similar phrasing, in the case of ML algorithms. What we will not do is buy into the argument 'we are all biased, therefore all our algorithms and computer software are biased'.

7.6 Computer Program Inadequacy

Some computer programs are unreliable. That will not come as news to you. Some unreliable computer programs are used in circumstances where their output, advice, or decisions have serious human or material consequences. A regularly cited example is risk assessment in the criminal justice system (Tashea 2017; Angwin and Larson 2016; Budds, Budds, and Budds 2017). In some jurisdictions software is used to predict whether convicted felons are at risk of offending again. This kind of software can

make false positive errors in classification (that a felon is at risk of reoffending, when the felon actually is not) and false negative errors (that a felon is not at risk of reoffending, when the felon actually is). (See Appendix C.1 for further explanation of false positives and false negatives.) Certainly, some examples of this software seem to be very poor. There is also a more general concern in this setting and that is: in a court of law the purported evidence should be transparent and contestable. The parties should know what it is and be able to argue about it. But many ML systems can lack transparency and not be revealing about their inner mechanisms (Liu 2019; Abebe et al. 2022). There are also many other examples of poor, and potentially damaging, classification software (e.g. credit ratings, job application assessments, mortgage lending decisions) (O'Neil 2016).

Computer software does not have to be unreliable. Some software can be proven to be correct— there are mathematical proofs that the software meets its specification. Other programs can be validated and evidence, and certification, provided that they meet requirements. There is a considerable portion of software engineering given to correctness and assurances of performance in mission critical settings. The development of programming languages and programming techniques has been in part driven by the need to produce quality in the face of complexity. Any computer science graduate will have had exposure to questions of how to ensure that a program is correct and how to produce evidence that it is. Just to give a couple of examples:- there is Unit Testing, where test code is written at the same time as the actual programming and run automatically over and over as the program develops, and Extreme Programming, which emphasizes teamwork in the actual programming. That said, there still can be

unreliability in the end result, and, in the case of ML, there is another factor. Many of the programs are quasi-empirical.

ML and DL can often be more akin to empirical science than they are to traditional computer science and the practices of software engineering. What is being asserted here? In general, our knowledge can be divided into empirical knowledge and non-empirical knowledge. Empirical knowledge is knowledge assured by observation and experience. One form of empirical knowledge is that provided by science. Scientific method includes deliberate experimentation, random controlled trials, natural experiments, and the like. Science is conjectural and fallible— there is the permanent possibility of error. It is also, for the most part, implicitly or explicitly, probabilistic (Howson and Urbach 2006). The theories, explanations, and predictions involve probabilities. In contrast, non-empirical knowledge, for example, mathematics, is knowledge assured by logic and reason. It is not usually conjectural, fallible, or probabilistic.

Computer programs, their correctness, and our knowledge of what they do, are generally in the domain of the non-empirical. (There are exceptions such as non-deterministic algorithms, genetic algorithms, and so forth, but these are but small paddocks in the large continent of computer science.) ML and, especially, DL, is another matter altogether. It is on the empirical side of the divide— it is quasi-empirical— and, also, most of its predictions are probabilistic.

Many DL systems are black boxes— quite how they work internally in specific implementations is often unclear. Then whether they actually do

work as anticipated is often a matter of experiment and empirical test. Evidence is gathered by providing the systems with data and seeing if they behave as they should. The testing is made more complicated by the probabilities simply because any probability other than 0 or 1 is consistent with any actual outcome in the world. For example, if risk assessment software predicts that a specific felon has an 80% chance of re-offending, that prediction is *not* refuted by the felon *not* re-offending. (Just as, if the weatherman does not have to be wrong by saying that there is an 80% chance of rain and then, actually, in reality, it turns out that there is no rain on the day in question.) Science itself is empirical, and it has evolved techniques for dealing with the probabilities. So, DL is not beyond redemption here— it is just that there are challenges and deeply entrenched fallibility. Humbleness is the order of the day.

7.7 Bias in the Context of Wider Machine Learning Programs

Let us first consider what ML can and cannot do to address unfairness and representation in a general setting. This is important for librarians to know. Librarians offer education in 'information literacy'. Knowledge of the properties of ML will increasingly become a core aspect of this. Then we will look at ML unfairness, representation, and classification, specifically in the case of librarianship.

7.7.1 Fairness ('Distributive Justice')

> The economic, political, and social frameworks that each society has—its laws, institutions, policies, etc.—result in different distributions of benefits and burdens across members of the society. These frameworks are the result of human political processes and they constantly change both across societies and within societies over time. The structure of these frameworks is important because the distributions of benefits and burdens resulting from them fundamentally affect people's lives. Arguments about which frameworks and/or resulting distributions are morally preferable constitute the topic of distributive justice (Lamont and Favor 2017).

ML has little or nothing to add to the vast and supremely important topic, or concern, of being fair, the topic of distributive justice. ML is statistics concerning facts, it does not offer moral guidance.

However, ML can itself supply facts that allow decisions to be made. Also, research in ML has also produced some surprising results concerning fairness (e.g. some obvious strategies for being fair can harm the folk they are trying to be fair to).

We need some background to introduce ML into a discussion of being fair and unfair. Assume a classification is going to be done into two classes on the basis of a single numerical score (see, for example, (Wattenberg, Viégas, and Hardt 2022) and (Hardt, Price, and Srebro 2016)). These classes are used to make a prediction for entities that are being classified, and resulting prediction is either correct or incorrect. An example might be a judgement on whether a person will pay back a mortgage, using 'credit-worthiness' as the numerical score. We have information, or data, on the past scores, and

we also have information, or data, on whether the borrowers paid their mortgages back. The data may be like this:

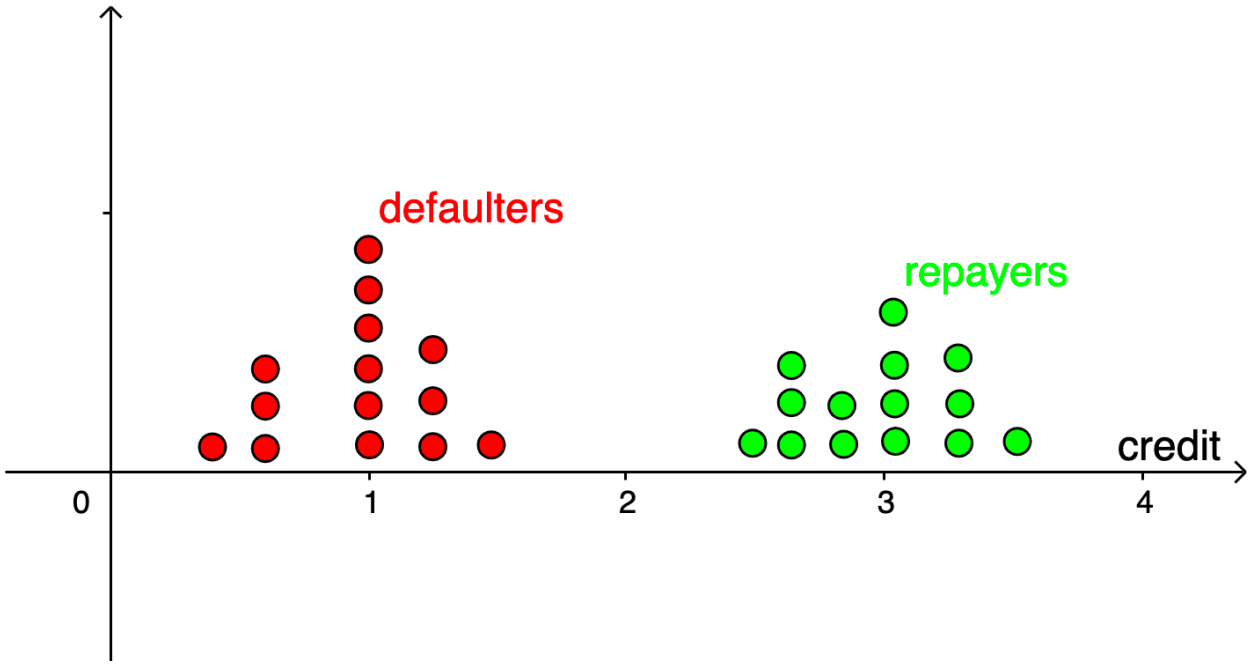

Figure 20. Graphic depicting defaulters and repayers against 'credit-worthiness'.

The intention here is to depict that anyone represented by a red dot (i.e. with a credit score between 0.4 and 1.5) failed to repay their mortgage, whereas everyone represented with a green dot (i.e. with a credit score between 2.4 and 3.6) did repay their mortgage. It is easy in a case like this to put in a credit score that classifies or divides the borrowers into defaulters and repayers— having a credit score of 2, or above, will do it

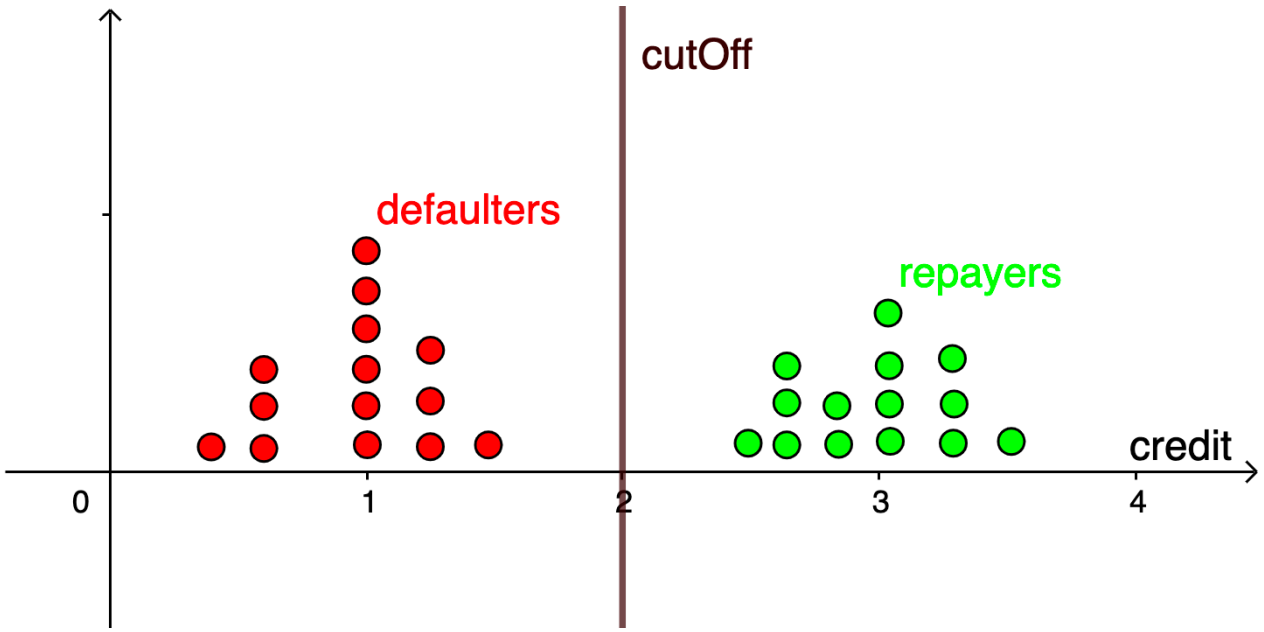

Figure 21. Graphic depicting defaulters and repayers against 'credit-worthiness' with a cut-off line.

This putting in of a cut-off line— a 'threshold classifier'— is a 'theory' that works perfectly on past data, and we will assume that it holds good with future data going forward. It is a theory that may have been devised by ML, or it may have been produced in many other ways. (After all, mortgage companies had similar theories long before the arrival of ML.) Problems with the theory start to arise if in the actual data these two regions overlap, say

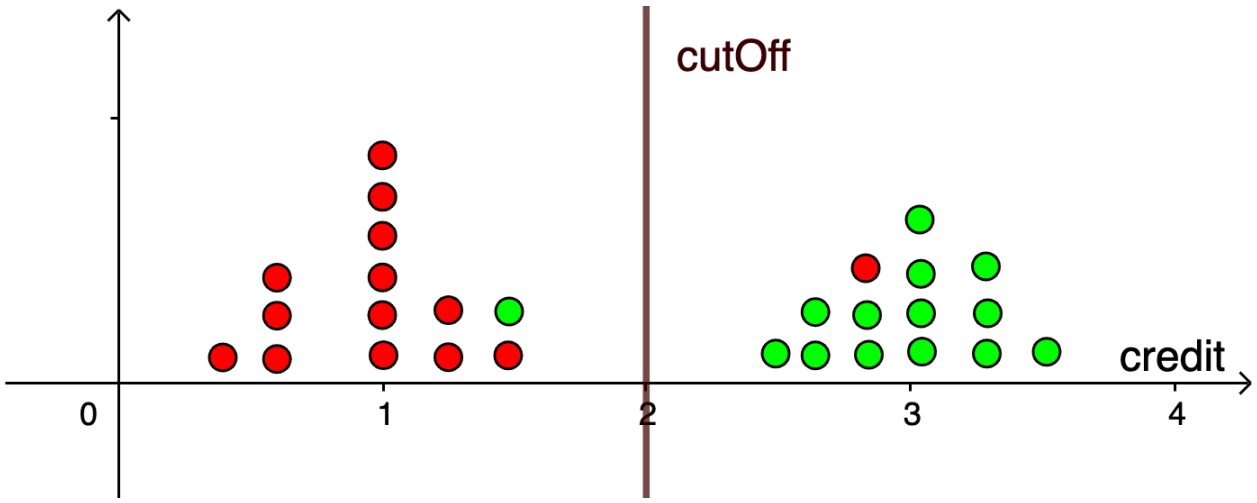

Figure 22. Graphic depicting defaulters and repayers against 'credit-worthiness', where there are false positives and false negatives.

With the data depicted in Figure 22, there is no way of putting in a classification boundary that does not make some errors. There are two kinds of errors that might be made in respect of repayment: false positives and false negatives. The false positives are where the classification suggests that borrower will repay, but the borrower does not. The false negatives are where the classification suggests that borrower will not repay, but, actually, the borrower does repay. For example, say a boundary was at credit 2.0, as in Figure 22, there is one false positive and one false negative. Of course, in a real case there might be thousands of borrowers, and hundreds of false positives and false negatives. No matter where the cut-off line is put there will be errors of these kinds. It is inevitable with this data. Statistics has techniques for adjusting theories to minimize errors. We will not invoke those here. Instead, we will leave this part of the discussion noting that

likely there will be false positives and false negatives. (Please see Appendix C for a further explanation of false positives and false negatives.)

That there are going to be errors is not a good result. Maybe the composition of the single score could be improved so as to separate the classes. This single number credit-worthiness score would typically be an amalgam of many other numbers, i.e. features, for example, salary, number of years of employment, family size, etc., adjusted with weights to reflect their importance. A DL approach would typically add features (any features not required would just end up with weight zero). This might help. But there will always be other general worries about the training data, whether the sampling properly represents the target population, whether the data on the features and the predictions (the labels) are correct, and so forth. A prudent conclusion might be that there will always be errors.

There is another aspect to what we know and what we do not know here. We know only probabilities. We do not know of a particular future borrower whether that individual borrower will repay. We know, for example, of borrowers like that borrower (i.e. ones with the same credit score) that, say, there is a probability of 90% that they will repay.

Let us move on from a factual judgement (on who is going to repay) to policy. Just as far as the actual lending goes, the mortgage company does not have to act inexorably on the advice or prediction of its repayment theory. It may have other reasons for lending or not lending. For example, it may have only so many funds to disburse and thus be forced into not lending to some clients whom the company knows would be perfectly good

repayers (or would have a very high probability of repaying). Let us now understand the cut-off in a slightly different way, as a major input to policy regarding future applications— that it is a main factor in separating those suitable for a loan from those not suitable.

The boundary of suitability can be adjusted to alter the proportions of false positive 'suitables' or false negative 'unsuitables'. For example, the mortgage company could potentially loan to every applicant (i.e. the cut-off would be at a credit score of zero); then there would be no false negatives (but, presumably, a number of false positives). Or, it could move the cut-off far to the right and have no false positives (but a number of false negatives). The mortgage company has choices. There are going to be errors. But the kinds and numbers of errors can be manipulated.

Now let us introduce fairness. Among the features in the single number credit-worthiness score there may be values for attributes, for categories, that would merit scrutiny in the context of bias and fairness, for example categories like gender and race. In this area of research, these categories are known as *protected* or *sensitive* features or categories.

Let us introduce a fictitious sensitive category, Shape, which has two values {circle, cross}. We can show these on our diagram.

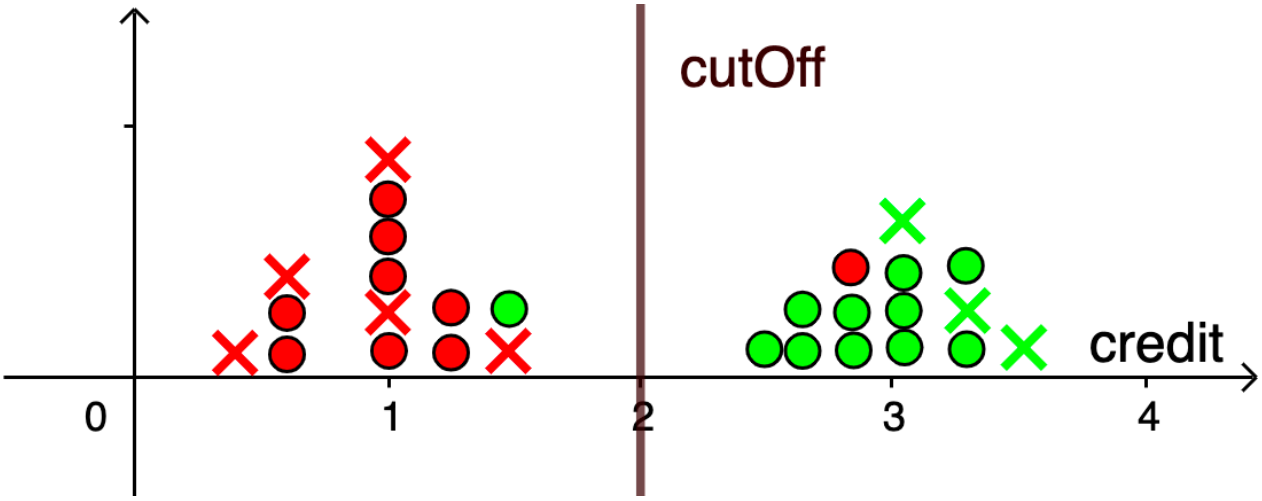

Figure 23. Graphic depicting defaulters and repayers against 'credit-worthiness', where there are false positives and false negatives and protected features.

Is our theory unfair on Shape? What might 'bias' mean in this context? How might we counteract bias? For the moment let us set aside the question of whether shape, e.g. being a cross, may cause a borrower to be a better or worse repayer. We will come back to that.

Here are some suggestions (Hardt, Price, and Srebro 2016; Corbett-Davies and Goel 2018; Kusner et al. 2018)

> Approach 1 ('unawareness' or 'anti-classification'): Ignore the property of Shape. The suggestion is that it is unbiased, or not unfair, over Shape, in as much as it ignores Shape. Analogically, in a more realistic setting, the suggestion is that one way to avoid bias over Race and Gender is not to have, or not use, any data on race and gender. There are two problems with this, though. Shape (or Race or Gender) may be correlated with other features that have to potential to serve as proxies for the protected attributes. For example, maybe all the crosses live in one zip code, and the

circles in another zip code. Then a strategy of ignoring shape, but using zip code, may lead to, or reveal, unfairness on Shape even though the algorithm does not directly use data on Shape. It is better to have data on Shape and to prove that the algorithm produces fair results (using an acceptable definition of fairness and a technique for achieving it.) The second problem is that the protected attributes may indeed have a causal relation to the predicted outcome label. It may be that the crosses are better repayers than the circles, so omitting information on this may produce a theory, a cut-off, that is less accurate (leads to more false positives and false negatives). This may seem unlikely or implausible in a mortgage repayment example. But in a medical setting, it is known that there are many differences between the races and genders. Whites, in the US, are more prone to certain heart conditions (e.g. atrial fibrillation) than other races (Dewland et al. 2013). Sickle-cell anemia is predominantly a disease of those who live in sub-Saharan Africa (and descendants of earlier residents of that region) (Rees, Williams, and Gladwin 2010). Women, and not men, have reproductive systems for bearing children. Men, and not women, are exposed to the possibility of prostate cancer. Women live longer than men. It seems that information on race and gender would be useful in medical settings. The systems would need to be 'fair' in their uses of that information, but not using the information at all does not seem to be the right move. Ignoring sensitive attributes can lead to unfairness with those the system is trying to be fair to. Female violent felons have a lower rate of recidivism than do male violent felons. Omitting gender from recidivism calculations may lead to harming women.

Approach 2 ('demographic parity'): separate the data into two sets of data (two graphs)— one for circles, the other for crosses— then, potentially, use a different cut-off for each ensuring that the same proportion are candidates for loans. So, for example, if there are a thousand crosses, and the cut-off for crosses leads to 10% of them qualifying for loan, then set the cut-off for circles to ensure that 10% of them qualify (whether there be 100 circles or 10,000 circles). Some defense can be made of this, in certain circumstances. But consider the true negatives, on either graph. It may be that a mortgage company, following the demographic policy, has to lend to borrowers that they know will not repay. At

an extreme, say all the crosses repay and none of the circles do (i.e. repayment is causally related to a protected attribute). If 10% of the crosses are offered loans, then the demographic policy requires that 10% of the circles also be offered loans, even though they are not going to repay. This actually does not seem fair to anybody (or to the company).

Approach 3 ('equal opportunity'): again, separate the data into two sets of data into two graphs— one for circles, the other for crosses. And, again, there will be two cut-offs. But this time the cut-offs focus only on those who are judged to be repayers. Then, the cut-offs are set to ensure, in so far as it is possible, that the same proportion of circle repayers and the cross repayers are offered loans. If you are classed as a repayer, there is equal opportunity of being offered a loan, whether you are a cross or a circle. (Wattenberg, Viégas, and Hardt 2022) and (Hardt, Price, and Srebro 2016) favor this approach. Presumably there is the problem with it of false negatives. Say you are a repaying cross. Some of those are going to be incorrectly classified as non-repayers (i.e. they are false negatives). But once they are (wrongly) classified as a non-repayer they will not have an equal opportunity of anything. They are not going to be offered a loan and nor do they have a chance of being offered a loan. It may be that equal opportunity is fair for the group but not necessarily fair for every member of the group individually.

Approach 4 ('counterfactual fairness'): (Kusner et al. 2018) suggest the following. Work with individuals only. Then require, and prove, that the probability of getting a loan for any individual, who is actually a cross, is exactly the same as the probability of that same individual getting a loan, had that individual been a circle (i.e. counterfactually being a circle) and similarly in the other direction, from circles to crosses. That is, data on sensitive attributes is obtained and used. But it is used to show that the outcome results would be the same for all individuals even were the values of those sensitive attributes were different. This approach certainly plumbs a central intuition. For example, under it, with race, whether you are black or white does not matter as far as your probability of getting a loan is concerned. While counterfactual fairness is different conceptually to demographic parity, Rosenblatt and Witter have proved that the two lead to

equivalent outcomes (L. Rosenblatt and Witter 2022). Demographic parity is easier to work with.

To sum up. Fairness in ML algorithms is an active research area. There are a number of proposals. Most of them have merits and shortcomings. Since there are probabilities involved, with false positives and false negatives, it seems unlikely that any suggestion on fairness in ML, can, at one and the same time, be fair to all individuals, to all groups, and to all related parties. Most of the proposals can be proved mathematically to hold, or not hold, of the relevant ML systems. There can be evidence and accountability. Which theory of fairness should be used is not a matter for the ML programmers to decide. It belongs with distributive justice, and it is a decision for the wider constituents.

There is a take home here. If there is a test that has false positives and false negatives (and pretty much all real-world tests do— medical tests, driving license tests, law school admissions tests, etc.). And if some, many, or most of the false positives have some other property (say, having the race of 'green', or 'being a cross'). Or if some, many, or most of the false negatives have some other property besides testing negative (say, having the race of 'red', or 'being a circle'). None of that, by itself, means that there is any evidence whatsoever of unfairness. Further analysis is needed— further statistics, or further mathematics. Epidemiologists— one group with a knowledge and interest in these methodologies— have tools at their disposal. One is causal diagrams. The use of causal diagrams, with appropriate data, can provide (fallible) evidence for fairness or unfairness. [Causal diagrams are explained and discussed further in Appendix D.]

7.7.2 Debiasing Representation

> Man is to Computer Programmer as Woman is to Homemaker? (Bolukbasi et al. 2016)

That attention grabbing question, or phrase, is part of the title of an important paper by Tolga Bolukbasi and fellow authors. What they are alluding to is that natural languages have associations in them which reveal assumptions about gender stereotypes. Sometimes these assumptions are innocent, harmless, and possibly even useful, such as the association between being a Queen and being female. Often, though, associations between words can be suspect and perhaps even revealing of undesirable underlying biases, such as that between 'receptionist' and 'female'. A problem in the context of ML is that if ML uses natural language as data, and it often does, the resulting ML programs might entrench or even amplify the biases. While this Bolukbasi paper focusses on gender stereotypes, it also would have application with racial or religious or other stereotypes. (See also (Caliskan, Bryson, and Narayanan 2017).)

This type of bias is different to the unfairness biases of allocation, for example, as to who gets mortgages and who does not. Rather, this is to do with biases of representation, with natural language processing (NLP) and with removing unwelcome stereotypical associations. Natural languages change, of course, and unwelcome stereotypical associations come and go. How to interact with that in a positive way is a larger question. But reducing bias, or 'debiasing', text which is used as input data to ML is a distinct possibility. The Bolukbasi paper has definite sound proposals on

this. It is not being asserted here that text for ML can be 'purified' perfectly. However, the text can be improved, and it should be possible to ensure that ML programs do not amplify existing biases in language.

There can be other harms of representation in addition to those strictly in NLP. For example, there can be such harms with the labeling of images—the attachment of metadata to images. In the case of the single sentence characterization of an image there might be denying people the opportunity to self-identify, reifying social groups, stereotyping, erasing, and demeaning (and probably further types) (Wang et al. 2022).

7.7.3 Panopticon Bias, the Panopticon Gaze

Certainly computers, artificial intelligence, and ML are enabling surveillance as never before. Examples of this are readily available in librarianship. There are recommender systems which can recommend books, articles, music, films, etc. that individual patrons might like. But to do this, the systems have to know what at least some patrons have read or explored in the past. Likely, patrons will have to give up some privacy to get the value-added intermediation of recommender systems. A more extreme example is that facial recognition software could track everything that every patron does in a physical library. This would be completely against the ethos of librarianship.

Facial recognition technology certain raises questions. It is a technology that allows the identification and tracking of individuals. These days it is

pretty good in a technical sense i.e. good at identifying and tracking. But, Nick Thieme asserts:

> AI's unique talent for finding patterns has only perpetuated our legal system's history of discrimination… Since people of color are more likely to be stopped by police, more likely to be convicted by juries, and more likely to receive long sentences from human judges, the shared features identified are often race or proxies for race. Here, computational injustice codifies social injustice. (Thieme 2018)

Joy Buolamwini has written on topics related to this. One of her early papers observes that she— a person of color— was largely invisible to computer systems, then later she offers the view that computer facial recognition was a technology of discrimination against people of color (Buolamwini 2019; 2016; *Race, Technology, and Algorithmic Bias* 2019). She has a new 2023 book *Unmasking AI: My Mission to Protect What is Human in the World of Machines* (Buolamwini 2023) The American Library Association also have a piece, now mildly dated (American Library Association 2018). We can all agree that recognition and tracking is a creepy technology that seemingly we can do without.

Or can we? There are many occasions when there is a need to know a person's identity— i.e. who the person is. In librarianship, there is the need to know who the patron is that is checking out the books. To establish identity there needs to be some gold standard, some difficult to forge validator whose original is on file or permanent record somewhere. Biometrics offers a way in here: it can use images of faces, fingerprints, images of irises, DNA, and similar. Right now, facial images are by far the

best combination of ubiquity and convenience. More-or-less everyone in the US has ID (identification) and that ID is going to be a driving license, a Real ID equivalent, a passport, a Green Card, or similar. All of these carry an image, a photo, identifying the holder of the ID. We can add to this folk who unlock their smartphones using a scan of their face. Facial recognition itself is now so good that it can recognize a person, in person, from a suitable image with 99% or more accuracy (quite what this 99% figure means is another question). Let us insert an anecdote. In June, 2023, the author flew from Dallas to Paris on American Airlines. When boarding he walked straight on to the aircraft in seconds, being identified by facial recognition (the airline, along with many others, already had a scan of his passport). Now, this facial recognition presumably was being trialed and not mandatory. But, also presumably, objectors would have had to produce their passports, to have printed and produced their boarding passes, and to have spent minutes with these processes. There will be no need to make facial recognition mandatory for these kinds of circumstances. We will all want it, for convenience. We will be falling over ourselves to get it. [Hot off the press, *The Independent* headline 7/18/2023 'Eurostar passengers leaving London can skip passport checks with new facial recognition tech'.] Separately, more than a few sporting venues use facial recognition technology to identify season ticket holders and to admit them without fuss or muss (Gee 2023). The US Immigration and Naturalization Service use facial recognition at airports to identify persons of interest. The author has been through INS at US airports many times. Every time until recently the INS agent took his fingerprints. On coming back from Paris in 2023 this did not happen. The agent told him that it was no longer necessary. Who knows why? INS must have had all the identification they needed. Perhaps from

facial recognition? It is hard to see all this being rolled back. Facial recognition has uses which are absolute winners. Of course, tracking people 24 hours a day, 7 days a week is an entirely different matter. There are companies (e.g. IBM) who say they will not sell this technology to the police (Peters 2020). There is an important difference between being recognized through a driving license in a pocket and being identified by facial recognition. Ordinarily, an officer of the law, or similar, would have to ask to see a driving license in a pocket, and the person asked might consent or refuse to reveal it. Then this kind of transaction would not scale, say to 10,000 people in a crowd at a protest. Facial recognition, though, can work with or without assent and it scales easily to many thousands of faces.

7.7.4 Bias in (Librarianship) Classification

This topic is included in Chapter 7 Machine Learning Bias and Librarianship.

7.8 Stochastic Psittacosis: LLMs and Foundation Models

The three reports or papers *On the Opportunities and Risks of Foundation Models* (Bommasani et al. 2022), *Language Models are Few-Shot Learners* (Brown et al. 2020), and *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* (Bender et al. 2021) have a wealth of material on potential harms arising from Large Language Models or Foundation Models. We need to be aware of some of the problems.

Large Language Models are large, no surprises there. Then what they do, essentially is cloze tasks (see Section 3.5, for an explanation of cloze tasks). This can lead to many other abilities (question answering, chatting, reasoning etc.). Nevertheless, what seems to be happening is probabilistic symbol manipulation on a grand scale. There is no doubt that some of these systems would pass the Turing Test (which at one point in time was taken to be an indicator of whether a system had intelligence (Oppy and Dowe 2021)). Nowadays received opinion is that the Turing Test is not demanding enough. There is the open question of whether there is emergence here. One view is that as the cloze tests, and the models mastering them, get ever more elaborate, something 'emerges' from the complexity, and perhaps that emergent property is true intelligence or even consciousness or sentience. A contrary view is that we can easily fool ourselves over apparently intelligent behavior. Maybe even 50% of the population would think that the ELIZA chatbot is either a real person or some truly intelligent software. The contrarians would just view the LLMs as being more complex symbol manipulators. This debate matters in the following way. We need to be cautious and skeptical towards what these modern sophisticated models have to say and recommend. We do not really know how they work in detail. Nor do we know, or can explain, the reasoning that produces many of the specific results. They have characteristics of a black box or oracle. Viewing them as stochastic parrots— to use (Bender et al. 2021)'s delightful label— lessens our awe in what they seem able to do.

One potential harm is misuse. GPT-3 can write, say English, as well as a native educated speaker. This opens some unwelcome possibilities:

> Any socially harmful activity that relies on generating text could be augmented by powerful language models. Examples include misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting. Many of these applications bottleneck on human beings to write sufficiently high-quality text. Language models that produce high quality text generation could lower existing barriers to carrying out these activities and increase their efficacy (Brown et al. 2020).

Basically, there is very little that can be done about this. It may become possible for other AI applications to recognize, and maybe filter, machine written text. If the writer application always puts in a 'watermark' (some giveaway combinations or sequences of words), it might be easy to recognize AI generated text. Using a machine to generate text is not always bad. For example, summaries of text, journal articles, and even entire document collections, can be exactly what readers require.

Another potential harm is bias. Bias is a huge and complex topic (as we have been seeing). Just briefly here we will use Kate Crawford's distinction between harms of allocation, harms of representation, and harms of classification (Crawford 2017). Harms of allocation, e.g. fairness in the availability of mortgages, should be able to be addressed. Within the bounds of sloth and fallibility, unfairness can be detected and remedied. Harms of representation, e.g. Muslim-violence bias (Abid, Farooqi, and Zou 2021), are a much harder case. There are hundreds of training data sets of (English) samples. The ones of these that are large and containing substantial source material from the Internet (e.g. from Common Crawl) will have some biased content. Then, if the training data has bias then so

too will GPT-4 and similar large models. To address this, there seem to be two central possibilities: to keep the unacceptable bias out of the training data, or to remove the bias from the model's output. Neither of these is promising. The systems use self-supervision on the training data precisely because it is near impossible to curate and label with the quantity involved. It may be that some software could filter the training data in some way. But that might not be entirely suitable. GPT-4, if it is going to have an all-around knowledge, needs to know about biased content (for example, holocaust denial). It 'just' needs to know that the biased content is biased and to not write biased content when writing in its own voice. Working on the output probably is not much better. Our experience with filters (e.g. using blacklists) on the Internet is not good. Filtering the word 'breasts', for example, to filter pornography tends also to filter 'breasts' in the context of breast cancer. It may be possible with systems like GPT-4 to instruct the systems themselves to remove biases of representation. For example, to provide examples of, say, anti-Muslim bias and prompt the machine to remove material like this from its reasoning and output. Harms of classification also are tricky and require attention. The actual classification categories used to classify people for some purpose, for example, usually or often depend on historical period, culture, and social factors (Hacking 1999). But training data for a foundation model will usually favor one time, place, and culture. This issue is also seen with medical data. There are many labeled sets of medical data. But many of them would not be good as training data. One reason is classification categories and diagnoses can change through time. Views of homosexuality in the 1950s, for example, are different from those of today.

7.9 Supplement: The Bias of Programmers

7.9.1 The 'Biases' of Professional Programmers

> … [the] possibility of programmer biases being encoded, in some way, in their programming artifacts.
> Contrary to making an error, which represents a single incident in which one makes an incorrect judgment, *a bias* is a systematic tendency to commit the same type of error over time or in different situations (Johansen, Pedersen, and Johansen 2021).

This kind of bias has absolutely nothing to do with bias in the sense of being unfair to anyone over race, gender, social status, economic status, etc.

A typical workflow process for a professional programmer at work on an individual program is that there will be a specification that the program must meet. There will be a 'house style', which is how the programs are to be written, for the employer, or for the open-source project, or for the intellectual area in question. When the program is first completed, and continuously thereafter, it will be subjected to quality assurance (including 'debugging'). The program will be shown to meet the specification. The programmer will have learned programming in MIT, Stanford, Princeton, Lomonosov State University in Moscow, or in many other worthy institutions including polytechnics and community colleges. They may even have learned from online sources or have been self-taught. They will have learned a *style* of programming. Not every programmer is different from every other programmer— definitely there are styles among programmers.

These are not 'biases'. All programmers make errors, but they will catch most of these in the debugging (and if they cannot, likely the program will not meet the specification). Most of the types of bugs are known. For example, there is an 'off-by-one-error' in a loop. (We need not worry about what this is.) Even the best programmers can make an off-by-one-error, but if a programmer has the systematic tendency to make off-by-one-errors (i.e. they are biased over this), basically they need to find a new line of work.

Johanna Johansen et al. argue that:

> … each program *probably* encodes the cultural and cognitive biases of their creators (Johansen, Pedersen, and Johansen 2021). [Emphasis added.]

and they offer evidence to this end. They even offer evidence of priming (that is, they can 'prime' a programmer to exhibit specific cultural and cognitive biases). Theirs is important work. But let us look at how it proceeds. There is a programming task— to write or sketch a program— but they omit any real specification for what the program is supposed to do. In this way they place the programmer in a condition of uncertainty. Then, their method suggests, the programmers reveal their cultural and cognitive biases in their decisions on what the program is to do, what the specification might be, (and this can be primed). While this research is important and addresses a rarely researched area, we find it unconvincing. Your mileage may vary. But our mileage is— if there are no specifications, all bets are off. There are starting to be other research papers in this area, and Johanna Johansen et al. provide a valuable guide to these (Johansen, Pedersen, and Johansen 2021).

7.9.2 The Biases of All of Us as Programmers

Pretty much all of us are programmers. There are six and a half billion owners of smartphones in the world ((Turner 2018) updated to 2022). That is over 80% of the world's population. Those phones likely have applications ('apps') on them, and settings. They are configured. That configuration, in conjunction with the host operating system and infrastructure, amount to a suite of programs, a cluster of software and software infrastructure. The owners of the phones, or their friends, or (young) relatives, or delegates will establish or program the configurations. There is a lot of programming going on. There will be extensive bias in the sense of systematic mistakes in the programming, and also the phone will enable bias, prejudice, in the sense of bringing to light unfair views or opinions about other human beings.

The importance of this is that to a large degree how our own smartphones, and similar, are set up are within our own control. We can reduce bias if we wish to. What we need is 'information literacy' and configuration agents to implement the prescriptions.

7.10 Annotated Readings for Chapter 7

Algorithmic Justice League. “Algorithmic Justice League - Unmasking AI Harms and Biases,” 2022. https://www.ajl.org/. (Algorithmic Justice League 2022)

American Library Association. “Facial Recognition.” Text. Tools, Publications & Resources, 2018. https://www.ala.org/tools/future/trends/facialrecognition. (American Library Association 2018)

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias.” Text/html. ProPublica. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Angwin and Larson 2016). This criticizes risk-assessment software in the case of Northpointe's software and Broward County, Florida. The article brings the area to life with real aberrant cases which presumably amount to false positives or false negatives (and which involve the sensitive feature of race). It links to a description of its own methodology and full data set. Northpointe do not agree with the analysis and its findings.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” arXiv, 2016. https://doi.org/10.48550/arXiv.1607.06520. (Bolukbasi et al. 2016)

Caliskan, Aylin. 2021. “Detecting and Mitigating Bias in Natural Language Processing.” Brookings (blog). May 10, 2021. https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/. (Caliskan 2021). Beautifully clear and comprehensive, and at about the right intellectual level for us.

Glusac, Elaine. “Your Face Is, or Will Be, Your Boarding Pass.” The New York Times, December 7, 2021, sec. Travel. https://www.nytimes.com/2021/12/07/travel/biometrics-airports-security.html. (Glusac 2021)

HAI. “AI Index Report 2023 – Artificial Intelligence Index,” 2023. https://aiindex.stanford.edu/report/. (HAI 2023). This is 386 pages long. You could try reading Top 10 Takeaways, which is 2 pages long.

Katell, Michael, Meg Young, Bernease Herman, Dharma Dailey, Aaron Tam, Vivian Guetler, Corinne Binz, Daniella Raz, and P. M. Krafft. “An Algorithmic Equity Toolkit for Technology Audits by Community Advocates and Activists.” arXiv, 2019. https://doi.org/10.48550/arXiv.1912.02943. (Katell et al. 2019)

Rainie, Lee, and Janna Anderson. “Code-Dependent: Pros and Cons of the Algorithm Age.” Pew Research Center: Internet, Science & Tech (blog), 2017. https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/. (Rainie and Anderson 2017). Much of the material in this Chapter has been on the 'Cons' of algorithms. But there are also some 'Pros'.

Wikipedia. “Algorithmic Bias.” In Wikipedia, 2022. https://en.wikipedia.org/w/index.php?title=Algorithmic_bias. (Wikipedia

2022b). This article, as it was 9/29/2022, is extremely good— both broad and deep. All the topics and discussion are important. In the present text, we are just not keen to apply the label 'algorithmic bias' to them. The article itself says 'In many cases, even within a single website or application, there is no single "algorithm" to examine, but a network of many interrelated programs and data inputs, even between users of the same service.' That captures where we are coming from.

Chapter 8: Bias in Machine Learning and Librarianship

8.1 Introduction

Let us start this Chapter by revisiting the paper by Su Lin Blodgett et al. on bias in NLP (Blodgett et al. 2020). The abstract to this, in full, is:

> We survey 146 papers analyzing "bias" in NLP systems, finding that their motivations are often vague, inconsistent, and lacking in normative reasoning, despite the fact that analyzing "bias" is an inherently normative process. We further find that these papers' proposed quantitative techniques for measuring or mitigating "bias" are poorly matched to their motivations and do not engage with the relevant literature outside of NLP. Based on these findings, we describe the beginnings of a path forward by proposing three recommendations that should guide work analyzing "bias" in NLP systems. These recommendations rest on a greater recognition of the relationships between language and social hierarchies, encouraging researchers and practitioners to articulate their conceptualizations of "bias"—i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements—and to center work around the lived experiences of members of communities affected by NLP systems, while interrogating and reimagining the power relations between technologists and such communities (Blodgett et al. 2020).

This paper is very thorough in its reasoning, evidence, and citations. Please read it. We will take from it its three recommendations, applying them to librarianship:

1. Recognize the relationships between language and social hierarchies.
2. Encourage researchers and practitioners to articulate their conceptualizations of 'bias'—i.e., what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements.
3. Center work around the lived experiences of members of communities affected by [the] systems.

Some material for the first recommendation is in (Blodgett et al. 2020) itself. The third recommendation requires extensive outside empirical research which we are not equipped for. That leaves as our focus:

> ...what kinds of system behaviors are harmful, in what ways, to whom, and why, as well as the normative reasoning underlying these statements.

There is a vast literature on bias in information provision (and, of course, information provision includes librarianship). But much of this literature has a wider ambit than librarianship. It wants, for example, to 'interrogate' internet companies, and power structures of one kind or another in society as a whole. No comment is passed on that here. We need focus on ML. There is an intersection of ML, bias, and librarianship, which we will explore shortly. But, as a summary of what is to come. There is a problem with data. ML needs data to learn from. But at least some potential data from traditional librarianship might have unwelcome aspects to it. For example, an ML program could easily learn the cataloging task of applying *Library of Congress Subject Headings* (LCSH) to new books and resources. But some fear that the LCSH labels themselves are suspect. So, according to

some, bringing ML into it would just be to 'reify White Supremacy' (Cordell 2020).

8.2 Harms of Omission

In Christianity, no doubt among other places, there is the distinction between sins of commission and sins of omission— the first is doing something that should not be done, the second is failing to do something that should be done. Similarly with harms associated with ML: some are harms of commission, others harms of omission. When Crawford et al. discuss bias in terms of harms of allocation, harms of representation, and harms of classification (surveillance), and Suresh and Guttag extend this out to 7 harms, these are all harms of commission (Crawford 2017; Suresh and Guttag 2021). But information providers, including librarians also have an interest in something different. Provision is an aim, or to use a word with a wider span: 'service'. Failure to provide service can be a harm. It would be a harm of omission. Harms of omission are a little harder to deal with at a methodological level than harms of commission. With a harm of commission, what or who did it— the causal agent— is available. Whereas with omission, what or who did not do it— the absent causal agent— often needs to be identified and may not be identifiable.

8.3 What to Digitize

As mentioned in Section 1.4, ML algorithms need input data in the form of computer digital text i.e. as structured sets of 0s and 1s. For text sources, or text corpuses, that are not born digital, this raises the question of what to

digitize. In some areas, the practice certainly has been to digitize primarily what might be characterized as being 'white', 'colonial' resources: White libraries, newspaper collections of White newspapers, and so forth. This is a tricky area. We know from the fate of the Google Book project that many do not want their resources digitized, or even the resources of others (see Chapter 1.2). Also, we know that many peoples, tribes, or indigenous peoples, do not want some of their cultural artefacts recorded at all, let alone digitized. In contrast, there is the argument that digitization selection can be an anti-racist action (see, for example, S.L. Zeigler, *Digitization Selection Criteria as Anti-Racist Action* (Ziegler 2019)).

Perhaps the ML research and practical initiatives can stay on the sidelines in this debate. As Elizabeth Lorang et.al. write, concerning the Library of Congress's role:

> ... the technology itself will not be the hardest part of this work. The hardest part will be the myriad challenges to undertaking this work in ways that are socially and culturally responsible, while also upholding responsibility to make the Library of Congress's materials available in timely and accessible ways (Lorang et al. 2020).

8.4 Search, Primarily Using Search Engines

Search is a filter. The searcher is initially faced with some pages, a site, a collection of sites, or the entire Internet, then search reduces the rich vista to something suitable for the occasion. The vista is filtered.

There are different possibilities here, and different possibilities for distortion or bias. If the Search algorithm uses keywords, spelling correction, semantic correction, stemming etc., various mistakes or manipulations can occur. Louis Rosenfeld et.al. report that an early instantiation of a search engine on Amazon responded to searches for the subject ‘abortion’ with the question ‘do you mean ‘adoption’?' (Rosenfeld, Morville, and Arango 2015). This suggestion is rather more than spelling correction. It is also regularly reported that various configurations of search engines misdirect searchers for 'abortion providers' to 'adoption agencies'. Search sometimes works via recommender techniques comparing the searcher and the searcher’s task to similar searches by other patrons. What happens here depends on which groups the search is compared with, and this can be, in some sense, fair or biased. Usually, a search returns a list of links in order of relevance. Now, relevance is topic, person, and occasion dependent (Frické 2012). It depends on the keywords, the person searching (different people may get different results from the same keywords, and the occasion (the same person may get different results from the same keywords on different occasions). The latter two features or aspects depend on the degree to which the engine is tracking the User (and the engine does need to be aware of previous searches in order to disambiguate, narrow, and help the User). There is also manipulation of various kinds. For example, there is Search Engine Optimization (SEO) which is tricking the engine algorithms to place some urls or links higher than they otherwise would be. It is known that most Users will not look beyond the first few links that are returned from a search. This might not matter for the supply of 'pure' information. But, for example, if you were a commercial entity, you would prefer to have the links to your products within those first few. The

provider of the links prefers this, not necessarily the User. There are companies that provide those services. The returns may thus be affected by paid interests, such as advertisers, retailers, or political groups (although most engines will identify paid links). The search engine companies try to identify the techniques of SEO and to immunize or neutralize them. It is a continuing battle. But even among algorithms that rank by genuine 'merit', there can be different, and sometimes equally acceptable but orthogonal, views of merit. [Orthogonality here means this. Consider the example of sports cars: is the one that goes faster better than the one whose looks are more head turning? Or vice-versa? It is hard to know. The two properties are orthogonal— they are independent one with the other.] Search is a filter infused with judgments and values (and orthogonality).

Some areas of the Internet are cesspits. That is one side effect of ease of content creation and freedom of speech. Safiya Noble, in her book *Algorithms of Oppression: How Search Engines Reinforce Racism,* reports that her 2009 search engine query for 'Black girls' returned the porn site 'HotBlackPussy.com' as its first hit. Later studies— 2011 onward— produced similar results for the search 'Black girls' (and also for searches for 'Asian girls', 'Asian Indian girls', 'Latina girls', 'White girls'). One conclusion that Noble draws is:

> ... girls' identities are commercialized, sexualized, or made curiosities within the gaze of the search engine. Women and girls do not fare well in Google Search—that is evident (Noble 2018).

These specific results have changed out of all recognition now, 2023 (possibly in part as a result of Noble's research).

What a search engine returns for a query does not depend solely on the query. It depends also on who is asking the query— that is what personalization does— and the occasion and history of the question being asked. Having the query 'black girls' (and similar) answered with links including porn sites certainly gets our attention. But what specifically is the harm here and who is being harmed? Some of the folks asking this kind of search question might be offended by this kind of answer (and some might not be, and some may even by pleased by it). Working to correct offense is tricky, which is not to say that it is not worth doing. The harm seems to be elsewhere, with girls, in general, and of having their identities commercialized, sexualized, or made curiosities. It is a harm of representation. Search engines have made adjustments in this area.

Some years ago, when the world wide web was just starting, there were curated pages of links on topics. These pages were authored by humans. For example, Lycos and Yahoo did this. These curated pages are like bibliographies, even annotated bibliographies, that a librarian might produce. They did not have or use keywords as lead-ins. Had there been a curated page on 'Black girls' it may well have been neutral as to range of topics, in so far as that is possible. Nowadays, there are just too many pages and too many links to make this practical. Certainly, there are 'pillar pages', or 'topic pages', or 'topic clusters' (see, for example, (Clariant Creative Agency 2022)). These are somewhat similar to the old, curated pages, but likely they will have been generated by computer program (possibly a ML program).

If, in a reference interview, a reference librarian was given the keyword string 'black girls', they almost certainly would think that it is under-specified. A librarian would ask for clarification and disambiguation. Search engines sometimes do that, with overly long keyword strings, but generally they are in the realm of guesswork. Once a search engine has an initial return click among first list of links, it can usually improve the suggestions. So, usually searching is a process, not a one-shot question and answer.

Also worthy of mention are autocomplete (autosuggestion), and 'trends'. Google, to take an example, will autocomplete a partial search string on the Google Chrome Browser. Here is what it did for the author on 10/27/2022 for 'black girls':

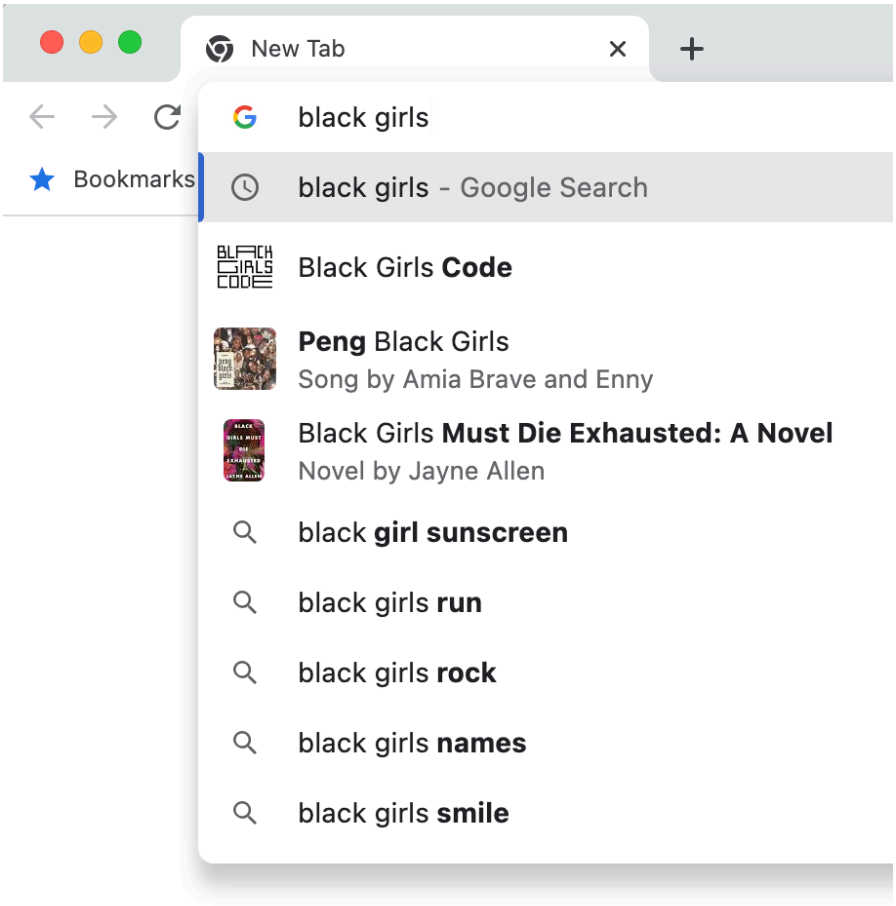

Figure 24. Screenshot of Autocomplete on the String 'black girls'.

We do not see anything untoward in that list. Google explains how autocomplete (and Trends) work on the page:

> https://support.google.com/websearch/answer/7368877?hl=en#zippy=%2Cwhere-autocomplete-predictions-come-from

In part, this reads:

> Autocomplete is a feature within Google Search that makes it faster to complete searches that you start to type. Our automated systems generate predictions that help people save time by

allowing them to quickly complete the search they already intended to do.
Where autocomplete predictions come from

- Autocomplete predictions reflect real searches that have been done on Google. To determine what predictions to show, our systems look for common queries that match what someone starts to enter into the search box but also consider:
- The language of the query
- The location a query is coming from
- Trending interest in a query
- Your past searches

These factors allow autocomplete to show the most helpful predictions that are unique to a particular location or time, such as for breaking news events.
In addition to full search predictions, Autocomplete may also predict individual words and phrases that are based on both real searches as well as word patterns found across the web.

Difference between autocomplete & Google Trends
Autocomplete is a time-saving but complex feature. It doesn't simply display the most common queries on a given topic. That's why it differs from and shouldn't be compared against Google Trends.
Google Trends is a tool for journalists and anyone else who wants to research the popularity of searches and search topics over time.

Notice here that autocomplete adapts to the User.

One general point to be made in all this is that the search engines are in competition one with another. They have incentives. In broad sweep terms, they need to be providing what the Users want, or need, or what is useful or valuable to them. Nowadays, Google has dominance with search engines. But that was not always so, and it does not have to be so going forward. There are various kinds of anonymous, non-tracking engines, and also straight-out competitors. In the absence of diktat, the Users will choose.

8.5 Social Media, Dis-, Mis- and False-Information

It is a jungle out there, and some ML programs have the potential to make the situation worse. ChatGPT can write English better than most English native speakers and writers. Essentially, readers would have difficulties in judging that ChatGPT English output had been written by a machine. ChatGPT could write disinformation tirelessly 24 hours a day 7 days a week. There is not much that librarians can do about this, apart from providing good education on information literacy. It may be that other ML programs, or even ChatGPT itself, could detect that samples of written English had been written by machine. This might be somewhat similar to plagiarism detection software. But being written by a machine does not of itself have to be bad. ChatGPT can abstract or summarize text. Summarizing today's newspapers, or this month's research journals, might be welcome and valuable.

8.6 Bias in the Organization of Information

8.6.1 Introduction

Traditional librarianship has devised such techniques as the ‘organization of information’ (content and container classification, abstracting, indexing, the use of surrogates, controlled vocabularies, thesauri, and the like) (Chan 2007; Rowley 2000; A. G. Taylor 2004).

There are questions of bias associated with these processes. To mention some:

- access to information, for individuals or groups, can be encouraged or discouraged
- straight out knowledge, or viewpoints, or theories, or beliefs or nexus of values can be conveyed or imparted
- attitudes towards the information resources themselves (including cognitive attitudes in the sense of social epistemology (Fallis 2006)) can be manipulated
- Mill's diversity of views (Mill 1869) can be promoted or obstructed
- Aristotle's diversity in the components of a good life (Wilburn 1999) can be promoted or obstructed

Potentially there are many topics that could be discussed here. We will restrict ourselves to a few.

[Appendix A has explanations of some slightly more technical librarianship terms that are used in Section 8.6:

1. emotive content: A.5
2. controlled vocabularies: A.2
3. classification and act of classification: A.6
4. taxonomies: A.1
5. thesaurus: A.1

]

8.6.2 Be Careful, and Sparing, with Emotive Content

The use of emotive content is valuable in advertising— it can help to sell things. It is also valuable in literature itself— it can manipulate our attitudes in liking or disliking characters and into becoming engaged with the story.

Librarians need to be careful with this, though. Any keywords, tags, subject labels, index terms, and so forth need to be low on emotive content. They need to be neutral. The reasons are: emotive content can produce harms of representation, and the use of emotive terms in indexes, say, or other stepping stones to content, will distort access.

ML systems should have little or no problem with avoiding emotive content. Large language models can just be prompted not to use emotive content. Some of the modern systems may be able to be given one example ('one shot') and then they will minimize emotive content.

8.6.3 Warrant and Controlled Vocabularies

A question is: where do the terms in a Controlled Vocabulary (CV) come from and what is the evidence or justification for introducing and using the particular terms that are used in a specific CV? There is a word or concept for this consideration. It is 'warrant' (Barité 2018). Warrant is important to us. It is one point where bias might arise.

There are different theories of warrant. We will briefly mention three: literary warrant, user warrant, and cultural warrant. Before explaining those, let us introduce a sample potential term for inclusion in a CV or some CVs : 'Gypsy Moth'. This is an example of a term that might have bias. Sabrina Imbler tells us in a 2021 article entitled *This Moth's Name Is a Slur. Scientists Won't Use It Anymore* that:

> The Entomological Society of America will no longer refer to common species of insects as "gypsy moths" and "gypsy ants," because their names are derogatory to the Romani people (Imbler 2021).

The article continues in part:

> For Ethel Brooks, a Romani scholar, the move is long overdue.
> As a child in New Hampshire, Dr. Brooks loved watching worms and caterpillars crawl across her hand. But one particular caterpillar, the hairy larvae of the species Lymantria dispar, terrified her. The larvae would swarm and strip the leaves from a tree, leaving behind so much destruction that people sometimes called them a "plague." But no one blamed L. dispar. Instead they blamed "gypsy moth caterpillars," the species' common name.
> "That's how they see us," Dr. Brooks remembered thinking as a child. "We eat things and destroy things around us."
> Dr. Brooks, now chair of the department of women's, gender and sexuality studies at Rutgers University in New Jersey, has spoken out against the use of the pejorative in fashion and college parades, she said. But Dr. Brooks never imagined the pejorative could be stricken from its use in the more staid realm of science.
> "It's hideous and super racist and it's hurtful," she said. "But what can you do about it?" (Imbler 2021).
> [To put some editorial interpretation here. Dr. Brooks is probably talking of the word 'Gypsy' as being 'super racist', not the phrase 'Gypsy Moth'.]

The name 'Gypsy Moth', or 'Gipsy Moth' (its British spelling) has not just been used for insects. Around 1930 the de Havilland aircraft company produced the Moth series of aircraft, which included the Gipsy Moth (and the related Tiger Moth, Puss Moth, Hawk Moth, Swallow Moth, Hornet Moth, etc.). When the supply of war-surplus eight-cylinder engines ran out, Geoffrey de Havilland designed the powerful and reliable four-cylinder Gipsy engine which was then used in the Gipsy Moth. The Prince of Wales owned a Gipsy Moth (you can see him portrayed flying it on the television series *The Crown*). Amy Johnson flew a Gipsy Moth single handed from Britain to Australia (roughly 11,000 miles at a flying speed of 100 mph).

> No other planes in their time and place so thoroughly served the advance of aviation (and civilization) in so many diverse ways as the de Havilland Moths (Harris 2002).

Separately, Gipsy Moth IV is the ketch that Sir Francis Chichester sailed single handedly around the world in 1967 (Chichester had worked earlier on the Gipsy Moth aircraft). It is reasonable to assume that Geoff de Havilland and Francis Chichester were not intending to demean their creations in any way by their choices of words 'Gipsy' or 'Gipsy Moth'. Quite the opposite, the names were intended to be positive descriptors. After all, de Havilland was selling mass produced airplanes. Then the items described, the engine, and the aircraft, turned out to be superlative. Additionally, Amy Johnson, the aviatrix, ' … was one of the most influential and inspirational women of the twentieth century (Gillies 2020)'. All was good for 'Gipsy Moth' in these domains, these periods, and these cultures. It is hard to imagine that any Romani were offended by 'Gipsy Moth' in the 1930s.

The idea of literary warrant comes from E. Wyndham Hulme at the beginning of the twentieth century. Lois Mai Chan et.al. summarize it

> ...the basis for classification is to be found in the actual published literature ... (Barité 2018; Chan, Richmond, and Svenonius 1985)

There is a slight difficulty or ambiguity here. What is 'actual published literature'? In 1911, that would have been published physical books that could be placed on shelves in libraries. But nowadays we have digital publication, e-Books, web pages, the Internet, and so forth. We need to take a wider view. Nevertheless, literary warrant would provide a motivation to using 'Gypsy Moth' in the 1930s, in the 2020s, and in the 2020s about historical 1930s literature, both in the sense of identifying an insect and identifying a training aircraft. Computers would be a great help here. The software does not necessarily have to be ML software. The processing has to look through the 'actual published literature' which it will be able to do really well (for all literature available in digital form). Notice here that with literary warrant there is one CV for all, whatever the area of focus is for that CV. The 'actual published literature' is the same for everybody. It is not one thing for one group of users, or culture, and another for a different group of users or culture.

User warrant focusses on the User. It is the User or the Patron (or, god forbid, the 'customer') that is trying to find the resources. Identifying the Users is not the easiest, nor is identifying how they do their searches. In the case of a public library, using say Dewey Decimal Classification, one might take the view that the Users are the 'public'. But it is also possible to be

more granular than this, patrons looking for materials on insects likely will be a different group to those looking for resources on historical aircraft. If the CV is for a smaller and more limited collection (e.g. for the index for a catalog for an aircraft museum), the tasks may be easier. Once again, computers are a great help. If patrons type into some kind of search box or Online Public Access Catalog or Discovery System, it will be easy to know what searches are being carried out. Nowadays most searches will be free text searches. The patrons can type in whatever they like and the software will learn what they do like. This time there can be several different CVs for the same collection of literature— one for one group of users of that literature and another for a different group of users of the same literature. Search software can personalize searches (as it would do for a private computer at home as opposed to in a public library).

Cultural warrant focusses on the cultures of groups of potential users of the resources. It is similar to user warrant in that there can be several different CVs for the same collection of literature— one for one culture of users of that literature and another for a different culture of users of the same literature. There can be obvious benefit in 'localizing' some CVs to place, time, beliefs, lifestyle, etc. i.e. to culture (no matter what literary warrant or user warrant might suggest in these cases). For example, legal systems can vary from state to state and country to country.

What are library practices here, and how can ML help (or hinder)? Typically, librarians will use universal systems, such as LCC, LCSH, DDC, to catalog, or provide metadata, for their resources. These CVs depend largely on literary warrant, sometimes with cultural adjustment. There is

widespread dissatisfaction with these CVs, particularly with cultural aspects. For example, Elizabeth Lorang et. al. write:

> Previous and ongoing collecting and description practices … were and are colonialist, racist, hetero- and gender- normative, and supremacist in other structural and systemic ways (Lorang et al. 2020).

Assume so. Were ML to use these practices going forward presumably it would just entrench them. (Mind you, if existing cataloging practices continue as is, without ML, there also presumably would be entrenchment.). ML does have the capabilities of correcting whatever shortcomings a CV might have. It certainly can reduce emotive content and provide cultural adjustment where required. Really some guidance is needed from librarians. For cataloging, librarians need to work themselves towards unbiased universal CVs (in so far as that is possible). ML can help with this (and also with the automation of insertion of values into the metadata fields of the resources).

What about 'Gypsy' and 'Gypsy Moth'? LCSH has this:

> Gypsy
>
> USE subject headings beginning with or qualified by the word Romani for topics related to the Romani people, e.g. Art, Romani; Romani poetry
>
> Gypsy moth [the insect, use as is]
>
> Gypsy Moth (Training plane)
>
> USE Moth (Training plane)
>
> (Library of Congress 2022)

This basically is literary warrant with some cultural correction. Notice, though, that using 'Moth' for 'Gypsy Moth' is asking for trouble with the precision of searches (because there are many other kinds of Moth training planes e.g. Tiger Moth etc.) Let us spell this out. You are interested in books on Gypsy Moth (Training plane), so you look up that subject in LCSH and learn that the catalogers have used Moth (Training Plane) for this topic, so you now search for Moth (Training Plane) and your search returns books on Tiger Moth, Puss Moth, Hawk Moth, Swallow Moth, Hornet Moth, Gypsy Moth etc. (most of which you do not wish to have). So, trying to be culturally sensitive, which we all want to be, has come, in this case, at the cost of ruining a search, and providing good searches is one thing librarians aspire to.

8.6.4 The Act of Classification Has Consequences

Obviously. In a court of law, classifying the accused as 'guilty' is rather different to classifying him as 'not guilty' (in terms of what the future might hold for him). So too for poisons, grades of scholarship applicants, and most everything else in daily life. (And, indeed, so too for being classified ‘black’ in (now fortunately historical) apartheid South Africa.)

Whole books have been written largely on this (cf. (Bowker and Star 2000)). Independently of the particular classification of items, we have noted how the accidental or intentional manipulation of the emotive content of equivalent concepts can affect attitudes to those things named.

Here are some of Sanford Berman's examples from the 1970s, and earlier, Library of Congress Subject Headings (S. Berman 1971):

> 'Yellow peril', 'Negroes, etc.', 'Mammies', 'Idiocy', 'Idiot asylums', 'Lunacy', 'Indians of North American, Civilization of,' 'Slaughtering and slaughter-houses—Jews', 'Barbarian invasions of Rome', 'Delinquent women'

Classification is also fallible, in common with all human cognitive endeavors. We know this directly from our everyday knowledge that anyone can make a mistake. We know it also from the test-retest unreliability of professional classifiers (catalogers) using precision and highly designed classification schemes (Snow 2017). If a classifier can classify the same item twice in two different mutually incompatible ways, classification is fallible.

The act of classification is also subject to bad-faith or 'rogue' cataloging. Everyone does this, all the time, pretty well every day. Faculty do it when they call a less than stellar student paper an 'A'. And some professionals do it too, fortunately not so often, perhaps when they have a private attitude, or strong feelings, which misdirect their work (and which, for example, might lead them to classify an information resource on abortion in such a way that the resource is difficult to find).

Classification is also subject to cultural factors. For example, different cultural groups have different attitudes to suicide, and thus different propensities to classify deaths as suicides, and this means that classification will not be isomorphic across cultures (independently of fallibility and rogue behavior).

In sum, the act of classification has consequences, can be used to produce attitude manipulation, is fallible, can be malfeasant, and has a dependency on culture. Some of the harms here are harms of representation, some are of allocation. Allocation can be seen as a harm of commission (doing allocation but not doing it fairly) or a harm of omission (failing to do allocation for certain groups or cultures).

8.6.5 Taxonomies Have Consequences

Taxonomies that have subclasses, i.e. most of them, make true or false assertions which are claims to knowledge. Asserting that one class is a subclass of another is a factual or conceptual piece of knowledge. For example, if a biological taxonomic scheme has whales as being a subclass of mammals, then it is offering the assertion that whales are mammals. Some of these schemes are intended to be objective i.e. they represent scientific or mathematical knowledge. Others are inter-subjective— for example, the Nurse Intervention Classification mentioned in Appendix A— and those represent decisions or conventions.

Some schemes can be problematic. Consider subjects, or topics, like 'creationism' and 'evolution' which we might be wanting to put into a hierarchical scheme of topics which allows them to inherit from some of 'scientific theory', 'false scientific theory', 'pseudo-scientific theory', 'religious view', and 'blasphemy'. What are we to put where? Well, who knows? Certainly, different groups of people would choose differently, and

different groups of people would disagree over which is a right or acceptable or appropriate classification. So, now any classification scheme does not so much contain knowledge, or, at least, uncontroversial knowledge, it does, however, contain or assert, a point of view (for example, one scheme might assert that creationism is a false scientific theory). There is a difference here between objective schemes and inter-subjective schemes. A mathematical classification that places the integers as a subclass of the reals is objectively right or wrong about that (and mathematicians can provide insight as to which that is). Similar considerations apply to the elements in Chemistry, or species in Biology.

Sanford Berman enlightens us that the 1970 Library of Congress Subject Headings are 'biased', and he means by this largely that they have unwelcome emotive content (S. Berman 1971). Berman also locates the bias. It lies with the classification schemes which are:

> …parochial, jingoistic Europeans and North Americans, white-hued, … Christian … heavily imbued with the transcendent, incomparable glory of Western civilization (S. Berman 1971)

Hope Olson echoes Berman telling us that the Headings reflect:

> …the exclusionary cultural supremacy of the mainstream patriarchal, Euro-settler culture' (Olson 2000).

And Bowker and Star wax long about generalizations of this: that the powerful subjugate the weak by imposing their will through classification (Bowker and Star 2000). (See also, (S. Berman 2000; Knowlton 2005; Olson 2002).)

Alright, but all schemes are 'biased'— biased in that they reflect a point of view (or knowledge, or beliefs, or opinions). We have to work with this. It is not an insurmountable problem. As Poincaré once remarked (about the indispensability of a point of view, when observing):

> He is no longer a slave who can choose his master (Poincaré 1905)

Then there are the conjectured conspiracies, surface or hidden, whereby the strong use classification to batter and imperialize the state of nature blissful. No comment on that here.

ML did not cause the problems in older library taxonomies (since they pre-date ML). However, care is needed to ensure that it does not prolong the issues. It may be used to counter-act them.

8.6.6 The Current State of Libraries and Their Organizational Systems

Melissa Adler writes:

> … libraries are complicit in privileging and circulating ignorance—inhibiting rather than opening up bodies of literature as sources of various knowledges (Adler 2017, 2).
> [Editorial note: 'Knowledges', plural, is an unusual form which perhaps has some currency in the sociology, or cultural aspects, of knowledge.]

Elizabeth Lorang and fellow authors write:

> Previous and ongoing collecting and description practices, for example, were and are colonialist, racist, hetero- and gender-normative, and supremacist in other structural and systemic ways. These understandings are the foundation on which training and validation data will be created and assembled; they will become reinscribed as statements of truth, even as we elsewhere champion the potential of computational approaches to uncover hidden histories, identities, and perspectives in collections. To engage machine learning in cultural heritage must mean confronting these histories, committing to the hard work of acknowledgment and rectification, and not simply reproducing them and giving them a whole new scale of power. There should not be a future for machine learning in digital libraries that is not first and foremost committed to, in the words of Thomas Padilla, "responsible operations" and to all of the ongoing, cross-cutting work that responsible operations entail. (Lorang et al. 2020)

The Thomas Padilla work alluded to is (Padilla 2019) *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* and in turn this cites influence from Rumman Chowdhury (Chowdhury 2023). Padilla writes:

> Chowdhury defines responsible operations as collective investments in, " . . . processes to combat algorithmic bias." (Padilla 2019 Note 6) [Seemingly, the original sources no longer exist.]

Padilla's Report is a research position paper for OCLC. It is substantial. We will just briefly mention the section on Managing Bias. Padilla writes:

> Bias management activities have precedent and are manifest in collection development, collection description, instruction, research support, and more. Of course, this is not an ahistorical frame of thinking. After all, many areas of library practice find themselves working to address harms their practices have posed,

> and continue to pose, for marginalized communities. As libraries seek to improve bias-management activities, progress will be continually limited by lack of diversity in staffing; monoculture cannot effectively manage bias. Diversity is not an option, it is an imperative (Padilla 2019, 9).

Then he suggests the holding of meetings, holding symposia, convening working groups, etc.

8.6.7 Designing Information Taxonomies for Librarianship

Classification and classification schemes are important in helping Users meet resources. Professional librarians, and similar, are *users* of classification, not *designers* of classification. Almost no librarians produce classification schemes, and, with the advent of copy cataloging, truly few librarians classify resources.

Here are some of the balls to be juggled. Classification is the most effective if it is in terms of the concepts, world-view, and values of the Users ('User warrant' as opposed to 'Author warrant'). But rarely are the Users a homogenous group. We all of us are simultaneously members of many different groups (male/female, old/young, Republican/Democrat, filmgoers, those interested in sports, gay/straight, etc.). And many of these groups truly define 'a culture'. We are all simultaneously members of many different cultures. Which culture, or cultures, is the classification designer aiming at? Presumably, something in the middle. A starting point is for the designer to wind down any emotive content in the scheme, and wind up the correct cognitive content. A quiet toned neutral, always fallible, cultural

absolutism would be the starting point. Less emotion and more knowledge generally facilitates access.

But designing is not so easy. Let us talk websites for a moment. Consider trying to design a website for the Flat Earth Society. Now, the classification, in terms of that culture, would reflect a point of view, indeed a somewhat extreme and false point of view, not knowledge. So, there is a switch here to cultural relativism. Folk not part of that culture, normal folk, might struggle with using the resulting website. But the site might not have been designed for them, and perhaps this does not matter. Normally, the duties of stewardship require that the Information Professional get the data right (for example, with Credit Card Bureaus). But a classification designer might have the need to get the data as the Users believe it to be, not as it is. The designer might also be asked to increase emotion or manipulate attitude, and not to be neutral. What would be wrong with the Luxury Jeweler's website classifying some of their diamond rings as 'Truly gorgeous', 'Prince Charming's choice', 'Sugar daddy's specials', etc.? But then there are cases like a White Supremacists' Web site. Here, let us guess, what the Users want is something expressing false views in an extreme fashion, possibly even hate speech.

The absolute limits seem to be freedom of speech. If what is envisaged is protected speech, ethically or legally protected speech, then the designer can instantiate it. He or she may have personal misgivings, in which case they should just not undertake a task of the kind that gives rise to worry. There can also be professional and societal misgivings that add caution to freedom of speech, as we will see. Of course, there are many ways

computers and websites can change attitudes, for example, there is ‘captology’— the persuasive use of computers (Fogg 2003).

8.7 Navigation: Metadata Supported and Otherwise

Conceptually, metadata for an individual information resource seems to be just a table of field-value pairs, where semantically the field-value pairs cover container metadata (such as Print Date, Physical Location (if these apply)) content metadata (such as Subject Matter) and mixed container-content metadata (such as Author, Title). All this seems to be entirely independent of navigation by a User from information resource to information resource, if such an operation is appropriate.

However, this is not quite right. Quite a lot turns on the nature of the values used in the field-value pairs. Obviously, one resource may have the same author as another resource and a patron may want to navigate to other books by the same author. Then dates, as values, have obvious relations to each other (e.g. earlier date, later date, same date). A similar point is true of locations. Then for metadata values for fields like subject matter, the values would likely come from a controlled vocabulary, possibly a thesaurus. Typical thesauri will support generic, instance, and partitive relations. In turn, these support navigation following these relations. For example, a book might have a metadata subject field value of 'rhopalocera' and a patron might want to navigate to another book, or books, on a similar but more general topic— of course, the topic is 'lepidoptera' and that topic will be a metadata field value on other books. The metadata is here supporting navigation, but it is doing so in conjunction with a thesaurus or

classification scheme. If that classification scheme is 'biased', the actual navigation might be distorted.

In the traditional case, using information resources which were paper pages, structured and co-located into books, placed on shelves, placed in libraries, placed in buildings, the mere physicality of the information resources enhanced certain styles of information resource to information resource navigation (for example, the patron browsed along a shelf from one structured information resource to the next). The use of surrogates (such as cards in card catalogs) liberated this to a degree (you could have several cards for each book and thus, using the indirection of the surrogate, you could make a book have as many near neighbors as you wished). When computers are used as a tool in classification and navigation, generalization is complete. A computer can easily transform views and presentations and it can easily provide different views of the same computer 'information resource locations'. Computers also can and will make free and extensive use of surrogates. For example, web links, or hyperlinks, are surrogates, and there can be thousands of links to the same one web page; and, in turn, those links can be organized and displayed in various ways. The designer can make it easy or difficult to navigate from information resource, or collection of information resources, to another information resource, or another collection of information resources. Navigation, browsing and reading navigation, is a filter.

Navigation and search pose similar problems. They can be used to manipulate access. The designer needs to be competent enough to know what might happen. Generally, from diversity considerations, one wants to

widen access. But there are many problematic cases discussed in librarian literature:- access to dangerous true material (dangerous to the User (e.g. 'how to commit suicide'), dangerous to employer (e.g. 'how to hack a web site and steal credit card numbers from it'), dangerous to society (e.g. 'reservoirs near you and how to poison millions of people using them'), access to inflammatory false material (eg Holocaust denial literature), and many other cases besides (Froelich 2004; Hauptman 1988; 2002; M. M. Smith 1997; Wolkoff 1996). Traditionally, librarians have been fairly 'hands off'. They argued that they have no assured knowledge as to the uses that an information resource will be put to, therefore, as policy, they can just supply information resources and not worry further. Nowadays, they are more cautious. ML should be able to help, no matter what the policy is.

8.8 Ethical Arguments to Underpin Assertions of Harms of Bias

There are ethical concerns that are the province of all informational professionals, such as: freedom of speech, freedom of access, privacy, intellectual property, stewardship, and the like. There are Codes of Ethics for various professional bodies associated with librarianship. Such codes usually build off distinctions between personal and professional ethics, duties to an employer, to the profession, and to society as a whole (American Library Association 2021; American Association of Law Libraries 2019; Society of American Archivists 2020; IFLA 2012). These can be a great help. More fundamental than any code would be the principles of non-maleficence ('do no harm'), autonomy ('let people choose for themselves'), informed consent ('give people the requisite information

for choice'), and perhaps even the Golden Rule ('treat others as you would like them to treat you'). And more fundamental still would be rights, duties, and ethical consequences.

8.9 Annotated Readings for Chapter 8

Bowker, Geoffrey C., and Susan Leigh Star. 2000. Sorting Things out: Classification and Its Consequences. Cambridge, MA: The MIT Press. (Bowker and Star 2000) Berman writes about this: "... the work is crippled by its own density and almost occult, inaccessible language. It is dizzyingly awash in definitions and theoretical formulations, too often stated in impenetrable infosci jargon (S. Berman 2000)." That said, the book has important material to offer on the power of naming, and the good and the bad that classification schemes can do. It has extensive examples from diseases, viruses, tuberculosis, race in Apartheid South Africa, and nursing.

Ziegler, S. L. "Digitization Selection Criteria as Anti-Racist Action." The Code4Lib Journal, no. 45 (2019). https://journal.code4lib.org/articles/14667. (Ziegler 2019). This has an extensive bibliography (with links to further bibliographies).

Chapter 9: What Might Natural Language Processing (NLP) Bring to Librarianship?

9.1 Introduction

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence. It is concerned with enabling computers to interact with natural languages such as English, French, or Chinese. It involves the use of algorithms and computational models to analyze and derive, and produce, meaning from human language in both written and spoken forms.

NLP made a huge step forward with the advent of Large Language Models (LLMs), such as:

1. Google's PaLM2 2023 Pathways Language Model 540 billion parameters (Narang and Chowdhery 2022). There is now AudioPaLM which is a large language model that can speak and listen (Rubenstein et al. 2023) There is PaLM-E for robots (Driess 2023).
2. Databricks's Dolly 2.0, 2023
3. Meta's LLaMA family (Large Language Model Meta AI), 2023
4. Microsoft's XLNet family (Generalized Autoregressive Pretraining for Natural Language Understanding), 2019
5. OpenAI's GPT-X family (Generative Pre-trained Transformer 3, 3.5, 4), 2020, 2023

Not all NLP relies on machine learning, or, indeed, on LLMs, but a good proportion does so.

While Natural Language Processing primarily addresses natural languages, quite often the resulting software can be applied to images, videos, or sounds. Such software is 'multi-modal'. This comes about because the underlying substrate is just numbers. Text is reduced to numbers, and so can images, for example. Then some of the algorithms are equally at ease with numbers sourced from text, numbers sourced from images, numbers sourced from sound recordings, numbers source from video, and so on.

NLP is becoming of increasing import to librarianship.

9.2 The Pre-Processing Pipeline

An NLP program typically will some pre-processing of the text prior to doing its actual task. Here are some of the steps that might be carried out:

- Format Normalization. This is the conversion of the text to a desired standard format (e.g. by making it all lowercase and removing redundant spaces or special characters).
- Word Tokenization. This will split the string or stream of characters into words or tokens. Usually this would involve looking for spaces, or periods, or separators between groups of characters.
- Base Form Normalization. Linguist often view words as consisting of a base form (a 'morpheme') plus, possibly, a prefix and/or a suffix.

For example, the word 'transportation' consists of (trans + port +ation). There are techniques to get the base form.

    - Stemming. This means 'reduced to their root form'. So, for example, 'consult', 'consults', 'consulting' would all be stemmed to the root form of 'consult'. The result of stemming does not always have to be a well-formed word; for example, 'change' and 'changing' would be stemmed to 'chang'.
    - Lemmatization. There is a similar technique to stemming— lemmatization— which does reduce tokens to a root word ('change' and 'changing' would be lemmatized to 'change').

- Parts Of Speech (POS) tagging. This identifies whether the individual tokens are nouns, or verbs, or adjectives, etc.
- There also can be the dropping or omitting of 'stopwords'. Words like 'a', 'an', 'the' etc. do not carry much information and so usually can be omitted.
- Dimension Reduction. More than a few times there are many more words or tokens in the source than are needed for the processing. For example, a novel might have more than 100,000 words in it, but an algorithm might need only 100 words, the right 100 words, to determine its genre. There are techniques for reducing the amount of input (and thus reducing the amount of time and the cost of the processing).
- In the case of LLMs, the tokens (the processing units that they work with) would often be larger than single characters but smaller than entire words.

These pre-processing techniques are also used occasionally within current librarianship. For example, in classical information retrieval one approach is to try to match a 'vector' of a query to 'vectors' of documents. With this, tokenization, stemming, dropping stopwords, and, possibly, a TF-IDF calculation, will have been done on the text of the documents to produce the 'vectors'. [To explain the TF-IDF calculation. A simple way of identifying which document, among many documents in a corpus, is relevant to a search is to consider how often a term appears. So, for example, a search for information about Toyota cars might look for how often the word 'Toyota' appears in the different documents. What is desirable is that 'Toyota' appears frequently in the document to be returned but not frequently, relatively speaking, in the other documents in the corpus. The calculation *Term Frequency Inverse Document Frequency (TF-IDF)* determines this.]

Going forward from the pre-processing, some of the later techniques will use a 'bag of words' approach. With this only the words (the tokens or lemmatized words) matter— what the words are and the number of times they occur. Other techniques will look at the grammatical structure of the text and perhaps try to parse some or all of it. Yet others might use text embeddings.

9.3 Text Embeddings and Similarity

We have already mentioned *word embeddings* in the context of Word2Vec— see Section 3.8 — and the problem there was to determine whether two different words had the same or related meanings. But there is

the similar, but more general, problem of determining the degree to which two entire text strings are related, or similar, to each other. For example, one of these text strings might be a search or query string that a User has entered and the other text string might be that of a complete document in a document collection— if these two are related we might theorize that perhaps the document has some relevance to the search. *Text embeddings* are one technique to address text string similarity. There are many different ways of producing text embeddings, and there are many different free and commercial software applications to do it (for example, Sent2Vec, FastText, Doc2Vec, or Gensim). Almost all of them will produce a vector (i.e. a list) of numbers to represent the strings and also provide a measure that reveals how similar two vectors are. So, as an example, OpenAI's text-embedding-ada-002 model embeds the string 'Stochastic Psittacosis' to the vector:

> [0.012791729532182217,-0.009504193440079689, -0.007625600788742304, -0.012044000439345837,-0.012828806415200233, 0.012532186694443226, -0.005901498254388571,0.003066616365686059, -0.00211805081926286 2, -0.0020809732377529144,-0.002658763900399208, 0.024434056133031845, 0.002084063133224845, -0.02558345906436443 ... where there are about 1500 further numbers in the list.]

The source texts strings can be of almost any length. Although it would be usual to split up long texts into chunks. So, for example, a text document could be split into chunks (into chapters, pages, paragraphs, sentences etc.) then vector for a search string could be matched into vectors for those chunks and the relevant chunks have a location or locations within the document. There are some cautions. Modern embeddings can be produced by LLMs. Some LLMs have been trained on data produced or written before

2020. This means that such LLMs might not perform embedding well on more recent text (text involving recent events, slang, or changed practices of speaking or writing). Then, once you are in the world of LLMs there might be bias. So, care is needed is needed to check LLM embeddings for bias. (There are tests to do this.)

Once there are vectors for strings, and a similarity measure, several further opportunities become available. The context here is a corpus of text documents which has been embedded. Additional infrastructure would include a vector store, which is a database of the embedded vectors.

9.3.1 Searching by Meaning (Semantic Search)

We are all familiar with 'Find', the string-in-string search tool which is close to universal in word processing, or text editing, software. What this does is to find occurrences of the target word, a keyword, say 'attorney' for example, in a document. But if 'attorney' is the search string it will not find 'lawyer' (i.e. a synonym). But, in contrast, embeddings can search by meaning. They search by similarity of vectors. The vector for 'attorney' (or for phrases in which it appears) will be similar the vectors for phrases with 'lawyer' in them. This is a very powerful addition to search. What it amounts to, roughly speaking, is search with thesaurus support built in. This, for example, allows for search by topic or subject matter. Also, similarity admits of degrees, so the results returned can be ranked by how similar they are to the query string. This is a type of ranking by relevance.

Semantic search can improve on keyword search even in cases that do not directly involve synonyms. Here is an example from (Fitch 2023). Consider a user interested in the topic 'the fall of John Major'. [John Major was a prominent British Prime Minister in the 1990s]. Were the patron to ask a librarian about this, the librarian would understand exactly what was being sought and likely would be able to find suitable material. But a keyword search simply would not work because, for example, there would be many different ways of saying 'the fall of' and, separately, many document sources with the words 'john' and 'major' in them. Searching by meaning should or would do better and that would be carried out by checking similarity of an embedding of 'the fall of John Major' to embeddings within the documents in the relevant document collection. There is a striking second example provided by Amr Kayid and Nils Reimers of Cohere (Kayid and Reimers 2022). They report that Elastisearch (a keyword search engine) returned to the query 'what is the capital of the United States?' an article on Capital Punishment. (This happened because of occurrences of the (key)words 'capital' and 'states'.) Semantic search would not make this kind of mistake.

9.3.2 Research Trails

Once there is matching of chunks to chunks, then research problems or topics can be followed automatically from chunk to chunk, document to document, far and wide. This also might be useful outside of pure academic research. As examples, it might help with legal cases or with settling patent priority disputes.

9.3.3 Classification

Embedding can support classification in different ways. It can facilitate clustering. For example, the task of dividing a document corpus into 20, say, clusters of similar documents can be carried out by dividing the accompanying vector database into 20 clusters of similar vectors. It can facilitate classification using a supplied classification schema such as a thesaurus, controlled vocabulary, or ontological vocabulary. It could do this by seeing which vectors in the vector database of the texts were similar to the embedding vectors of the schema vocabulary.

9.3.4 One Style of Recommendation

If some sample 'desirable' documents are supplied, from a User, a group of Users, or from an institution such as a library, then embedding provides an easy way of finding other documents (i.e. recommendations) similar to the provided ones and of ranking the recommendations.

9.3.5 Plagiarism Detection

The embedded vectors can reveal if two vectors are too similar (maybe even that they are identical). This perhaps would suggest that the one source is plagiarized from another.

9.4 Named Entity Recognition

Named Entity Recognition (NER) is the ability of the software or agent to recognize in a source text the references to particular entities, e.g. to London or to Sherlock Holmes or to the Korean War. NER can also usually classify the entities into people, places, institutions, etc. The references need not be in some canonical form. For example, one journal article might have the word 'London' in it and another article the phrase 'the capital of England'— these are two references to the same entity. NER using a large language model, such as BERT or a GPT-X, on news articles, research papers, social media posts, etc., is extremely accurate. As you would expect, it is not quite as good on new, or extremely rare, entities. However, it is plenty good enough for most applications and it is distinctly better than humans attempting the same task. There is a difference here between the NLP model essentially recognizing or creating the relevant ontology and ontological vocabulary for itself against being provided with an ontology and controlled vocabulary to work with. For example, chemists have an ontology of elements, and compounds, and NER could use that when analyzing research publications in chemistry. Accuracy is higher if the NER is provided with an ontology to use.

NER is valuable for librarians, as it allows a User to obtain, for example, all the articles on London in a collection.

9.5 Topic Modeling

Topic modeling can look at a text or collections of texts and determine their 'topics'. Then it can 'tag' those texts with their topics. There is a slightly older way of doing this, and a slightly newer way which would use LLMs and embeddings.

With the older style, topic modeling is done by unsupervised learning. So, the ML software is clustering the contents of the texts. In the general case it needs be told how many clusters to make, and it does not have any names for those clusters. It needs to be guided as to those also. The result is that various different parts of the same text, or various different texts, can appear in the same cluster, meaning these parts or texts are on the same topic as each other. This type of analysis could also be done on documents, on paragraphs, on sentences, or on parts of sentences. To explain this at the level of documents. There are some background assumptions here, and some common techniques. The assumptions are a) that each document is a mixture of topics b) that each topic is expressed using a certain vocabulary, i.e. using certain words, and c) that any word has a certain probability of belonging to the vocabulary for a specific topic. Then the calculation is merely a matter of counting the frequency of particular words (say the number of tokens of the word 'car') and seeing how the best clusters of words can be formed doing justice to the word clusters and the documents. Common techniques include Latent Semantic Analysis (LSA), Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA). (The word 'latent' here means 'hidden' or 'in the background'. See, for example, (Manning, Raghavan, and Schütze 2009) for an explanation of these

techniques. LDA can classify each document as a mixture of topics (Blei, Ng, and Jordan 2003).) This type of analysis is very quick. So, the analyst can certainly experiment by varying the number of topics to find a result that is acceptable.

Were modern style embeddings to be used, the general approach would be similar but the algorithms would use embedded vectors from a vector store.

The main use-cases of Topic Modeling are outside librarianship. As an example, security agencies may have an interest in which new threats are appearing in social media posts. Librarians, in their professional work, would typically have existing classification schemes that they would prefer to use.

9.6 Text Classification Problems

As we have explained under Section 9.3, certain kinds of text classification problems can be addressed using embeddings. But there are other techniques. There are a number of text classification problems that share a common ML approach. All that is needed are suitably labeled training data, produced in conjunction with the desired classification scheme, and then the ML system can classify the actual target data.

9.6.1 Shelving and Subject Classification

Text classification is similar to topic modeling, but it uses a supplied classification scheme. This means that the training of the ML system will be different. It will use supervised learning, and that means that there has to be labeled training data (which is always difficult to obtain). Modern systems are reasonably accurate.

Classification, especially shelving classification, is something different to subject classification, indexing, or CV tagging. With books, and Dewey or LC, for example, a book can only have one slot (this is because it can have only one position on a shelf). This requires that the classification is exclusive and exhaustive and that each book carries only one, perhaps complex, 'tag' (i.e. its call number).

9.6.2 Sentiment Analysis

Sentiment analysis, at its most basic, can pick out the emotive tone of some text. This might be useful in a commercial setting, for example, to process reviews of a product, dividing them into favorable, unfavorable, and indifferent. It typically works using only 'bag of words', i.e. the word tokens that appear in a document. It ignores what those words are saying. There are training data sets available for sentiment analysis. In that setting, supervised learning would be used to train an ML sentiment analyzer. You can see how a bag of words approach can go wrong with a naïve analyzer.

Say some example favorable words are 'good', 'brilliant', and 'excellent' and a text passage is written using negations only e.g. 'the product is neither good, nor brilliant, and definitely not excellent'. The review is unfavorable, but naïve sentiment analyzer might rate it as favorable because it contains only favorable words.

Modern large language models, such as BERT and GPT-X are very good at sentiment analysis. They would not use a bag of words approach. Also, they would not be trained using supervised learning. Likely they would be trained using self-supervision followed by reinforcement learning. Sentiment analysis might help librarians in any situation where they are trying to obtain feedback or opinions on items or policies or services or courses of action. The relevant patrons or constituents could be invited to provide their input, which then could be processed in whole or in part automatically using sentiment analysis.

9.6.3 Author or Genre Recognition

Most of the LLMs can work in 'one shot' or 'few shots' mode. The means that their user or programmer just needs to provide a few examples and then the LLM will be able to carry out the task that is illustrated. In turn, this means that merely by providing some examples of an author's work (or of books of a genre) and the LLM will be able to recognize which books are written by author (or of which books are of which genre).

9.7 Controlled Vocabularies, Thesauri, and Ontological Vocabularies

[Appendix A provides some background on the concepts used in this section.]

ML/NLP can create controlled vocabularies, thesauri, ontologies, and ontological vocabularies for any corpus of text. It would be able to do so swiftly. Semi-algorithmic techniques for producing these kinds of vocabularies are well known (see, for example, (Zeng 2005)). In recent years, computer support would certainly have been used for these tasks. But natural language processing adds power tools.

As a sketch of some of the techniques that might be involved. The terms in the documents can be clustered. This would be similar to topic modeling. These clusters can be arranged in hierarchies, and favored labels attached to each cluster. That would produce a rudimentary thesaurus and controlled vocabulary Possibly, Latent Semantic Analysis (LSA), Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) will be used. This will reveal the important concepts that are in use (and their relations). There could be named entity recognition (picking out, for example, that 'London' is the name of a thing or entity). There likely would be parts-of-speech (POS) tagging— identifying nouns, verbs, noun phrases, etc. The results would be a thesauri or ontologies better than those produced without ML techniques (and produce quickly).

Controlled vocabularies, and similar, are still important even with free text search and search by meaning. They can provide an interface, or commonality, between different documents or collections of documents.

9.8 Indexing and Automatic Indexing

A simple example of an index that we are all familiar with is that of an ordinary back-of-the-book index. The index itself consists of a list of *entries*. The entries are composed of *headings* and *locators*. For our purposes, we can assume the headings to constitute a (controlled vocabulary) thesaurus— so there might be sub-headings and sub-sub-headings, etc. in a hierarchical arrangement. Each entry is a pair of a thesaurus term and one or more locators. In the case of a book, a locator would just be a page number. So, the index itself has an easily navigable structure— the top-level terms are in alphabetical or filing order. But the book, in so far as its individual fine-grained topics are concerned, is a jumbled disordered mess. Were a User to be interested in finding out where the ship Pequod is discussed, described, or alluded to, in the novel Moby Dick, a proper index would make this task trivial whereas doing it by looking through the book would be a considerable challenge. An index provides simple, readily understandable, access to points of interest in an information mess.

As to creating an index in the first place, Bella Haas Weinberg writes (mainly under the assumption that the indexer is a human being):

> An indexer must be something of a prophet— envisioning the concepts likely to be sought by users of a document, expressing those concepts in terms likely to be sought by users, and providing cross-references from synonyms and alternative spellings as well

> as links to related terms to assist users in finding all the information that is relevant to their topics of interest (Weinberg, 2009).

Notice here that Weinberg refers to indexing in terms of concepts. It is possible to do *derived* indexing, which uses only terms that appear in the document or documents, or to do *assigned* indexing, which uses the concepts i.e it is indexing by meaning. Assigned indexing is superior (' ...to assist users in finding all the information that is relevant to their topics of interest '). Assigned indexing addresses the problem of synonyms and homographs, and also those very difficult cases where a concept is alluded to but not referred to explicitly. Derived indexing is pretty well trivial from a computing point of view (it is string in string searching). Assigned indexing is another matter entirely as it seems to require 'understanding' the material.

> 'Good indexing permits good retrieval'.

As the quote reveals, Weinberg sees the main intellectual challenge of indexing as being that of creating the controlled vocabulary thesaurus of headings that is to be employed. The secondary problem of scanning the text to catch all the locations for the locators in the entries is relatively routine and easy.

Indexes can be surprisingly sophisticated. Indexes typically work on structured sources or across collections of sources. So, for example, an index for a book might tell you that material relating to the concept 'machine' appears of page 37 and page 39. This is to conceive of the book as

being a structured source, with pages which are themselves sources, and the index is finding the component sources. Or, as another example, an index across a digital collection of full text sources might inform you that a particular topic is to be found within a few specific sources. In sum, indexes enable access to works within works: articles within periodicals; short stories, poems, or essays within a larger work; or individual papers from a conference.

Cumulative indexes would use the same thesaurus to index across several different sources; for example, across all the monthly issues for a year's publication of a journal.

AI/NLP can improve traditional indexing. It can do standard indexing faster than human indexers and the result can be of higher quality. It can produce a headings thesaurus without difficulty. Then fleshing out the entries with the locators is essentially trivial. Many of the actual cases of indexing will use an antecedently provided thesaurus. For example, indexing medical journals would do this. The complete task then comes under the heading of *automatic indexing*. Another advantage to AI/NLP indexing is that it can provide locators to outside sources, for example to other books or texts with references to Pequod (the whaling ship in Moby Dick). Such pathways to wider resources are valuable for researchers.

9.9 Abstracts, Extracts, Key Phrases, Keywords, and Summaries

Abstracts usually either give the content in a shortened, sometimes structured, form, or point to or indicate what treasures are to be found in the work or works but without actually providing those nuggets. Single sources can be abstracted, so too can multiple sources (as might be the case with abstracting the news from several different news sources into one abstract or abstracting many reviews of a source to produce an overall viewpoint).

Prior to the 1950s abstracting was always done by humans. But with the research of Luhn and others attention was given to the possibilities of automatic abstraction (Luhn 1958). This work is a sub-area of Computational Linguistics. The techniques were certainly computational, and some may be classified as AI, but, until very recently there was not really ML involved. Automation is important because there are many more documents that would benefit from being abstracted than there are human abstracters to do the work.

Abstracting would generally be either abstractive summaries or extractive summaries. The former tries to understand what is in the source and then paraphrasing it in a shortened form. The latter extracts and condenses what is there, without necessarily understanding it, using statistical features and 'signposts'. To illustrate the typical statistical features that might be used. Material at the beginning and end of the entire text is important, material at the beginnings and ends of paragraphs is important, words (other than

stopwords (i.e. ‘the’, ‘a’, etc.) that appear frequently are important and so on.

Researchers in this area do have techniques for assessing their results. In some form or other, these are normally comparisons against abstracts that have been produced by humans. (See, for example, ‘Rouge’ i.e. *Recall-Oriented Understudy for Gisting Evaluation* (Lin 2004).) It seems fair to say, as of 2021, the computational abstracts are *not* of higher quality than human authored ones. However, the software approach is thousands, millions, or billions of times faster than humans. So, for example, it can produce abstracts real time e.g. it can abstract news sources as fast as those sources are conveying news. There is Machine Learning in much of recent research on abstracting, but usually the ML is not used on its own but rather is augmented by tried and true methods (Widyassari et al. 2020). There does not seem to be any technical problem as to why ML abstracting would just not get better and better. Likely NLP abstracting will be the equal of, or superior to, human abstracting sometime in the early 2020s.

Producing Key Phrases or Keywords is a similar, but simpler, problem to abstracting. If it is desired that the ‘Keys’ appear verbatim in the text, then challenge is like extractive summary. Often, though, it would be preferred that Keywords come from a Controlled Vocabulary (CV). If so, maybe none of a text’s Keywords actually appear in the text. For example, the non-CV word ‘car’ might be in the text but the preferred CV Keyword term ‘automobile’ is not. ML is a possibility both with extractive Keywords and with pure abstractive Keywords (e.g. Controlled Vocabulary Keywords). Extractive Keywords can be approached with unsupervised learning

(Mishra 2021). This is a well-worked research area, not because the researchers are trying to help journal editors with lead-in Keywords for articles. Rather, the computational linguists want to extract important words for further processing. Abstractive Keywords (e.g. Controlled Vocabulary Keywords) could be set up as a supervised Classification problem. The CV would be the classes or categories, and the classification would note perhaps the best 8 terms that apply. These lead-ins specifically would help journal editors, and researchers and librarians looking for documents.

A summary of content would usually be longer and more detailed than an abstract of the same content. Summaries often rephrase the content. Otherwise, abstracts and summaries would be fairly similar. Summaries, or abstracts, may have specific features. They may have Named Entity Recognition. If the source is fiction, a summary may contain a plot outline or an identification of the plot type or story type. Summarizing across multiple sources is also a form of synthesizing or aggregating those sources— summarizing across ten different articles on linear algebra amounts to one way of aggregating those articles.

How might these techniques help librarianship? Summarizing documents may directly help Users (which may be human, or software) to find relevant content. Summaries may improve information access tools, such as catalogs and databases, by becoming part of the metadata that is used to characterize information resources. One of the first books written by machine learning— *Lithium-ion Batteries*— is a collective summary of 200 research articles (Writer 2019).

9.10 Text Mining and Question Answering

Text mining can extract potentially useful information from text such as information regarding named entities, topics, classification features, genres, summaries, and sentiments of reviews— techniques that have been described in this chapter. Any or all of these types of information may help a User narrow an information search and thus make it more precise.

Text mining is often used in conjunction with other techniques for example, with sentiment analysis, or with recommendation systems.

Later we will look at Undiscovered Public Knowledge (UPK)— that is knowledge that is in books or libraries that is 'undiscovered'. Text mining helps discover it.

9.11 Machine Translation

This involves automatically translating text from one language to another. The value of this to librarianship is obvious. These topics and techniques are discussed elsewhere in this book.

9.12 Evidence

It has been asserted in this chapter that NLP is capable of doing this-and-that. Researchers in this field do have evidence. There are benchmarks that

the systems are tested on (NLP-progress 2022). In fact, there is usually competition between the systems on how well they can do on the benchmarks. There are questions over the accuracy of the results of generative systems. There are difficulties with the prompts: with ‘noise’, context, and ambiguity. Also, an LLM will not usually predict the same thing twice— thanks to the probabilities. So, it not clear just what an accurate result is where an LLM is given the task of writing a paragraph in the style of Jane Austen. [There is further discussion in Section 4.8.1 on hallucinations.]

9.13 This Is Not Magic

Several of the topics and techniques mentioned in this chapter might sound a bit esoteric. But, actually, the means for a programmer (or library technical services department) to produce suitable software is simple and has been readily available since 2023. The best approach would be to use Large Language Models (LLMs), and to use one that has a public Application Programming Interface (API). For example, Open AI's GPT-4 does so. Then a programming environment like LangChain can be used to 'program' or configure LLMs from their APIs. The resulting program to perform one of the tasks mentioned in this chapter might have about 50 lines of code in it. A professional programmer might write 7 lines of code a day (when, as a pre-requisite, they have to think about and research the problem). So, the first program might take a week to produce. But the second and subsequent ones could be written in about an hour each. Go to it!

Since November 2023, there are have been even faster and easier ways of doing some of these projects. OpenAI has produce the building framework 'GPTs' which is a simple no-coding-required system for personalizing the multimodal GPT-4 Turbo to various NLP and LMM tasks (OpenAI 2023d). Neither programmers nor LangChain are needed for this.

9.14 Text Processing and Laws

Needless to say, there are laws and contracts that might restrict the unfettered use of NLP on document collections such as libraries. Further, librarians and their institutions can be intimidated at the mere prospect of legal cases (like the rest of us). The laws concern primarily copyright and intellectual property, and they differ from country to country. This area is complex. A good initial source is H. Andrés Izquierdo's *20 Artificial Intelligence and Text and Data Mining: Future Rules for Libraries?* (Izquierdo 2022).

At a rough handwaving level, we might say this. Copyright concerns the *expression of ideas*, not *the ideas themselves*. So, when Einstein wrote the theory of relativity, his actual words might have had some copyright protections but the theory of relativity itself did not. So, when NLP abstracts, or paraphrases, or summarizes, or text-data-mines, documents or collections *in its own words*, it might be that there would be no copyright concerns. In contrast, extractive abstracts or summaries, or quotation of passages verbatim, etc. might be problematic. Additionally, many copyright laws have exceptions for 'fair use' which might include use for research, teaching, and non-commercial uses.

Contracts are another matter. The owners of the intellectual property can seek whatever contracts they wish. Librarians, or their institutions, can agree, or not agree, to these contracts, as they wish. Some advice: librarians should agree only to contracts that are permissive on text-data-mining and other NLP techniques.

A further problem or issue is that more than a few times the owners of the intellectual property are unknown or untraceable (as might be the case with some historical documents).

9.15 Annotated Readings for Chapter 9

Izquierdo, H. Andrés. "20 Artificial Intelligence and Text and Data Mining: Future Rules for Libraries?" In *Navigating Copyright for Libraries*, edited by Jessica Coates, Victoria Owen, and Susan Reilly, 497–540. De Gruyter Saur, 2022. https://doi.org/10.1515/9783110732009-022. (Izquierdo 2022). Just scan this.

Jurafsky, Dan, and James H. Martin. "Speech and Language Processing," 2023. https://web.stanford.edu/~jurafsky/slp3/. (Dan Jurafsky and Martin 2023) This is a standard text. It is probably too advanced for us. The draft of the 3rd edition is available free on the web.

NLP-progress. "Tracking Progress in Natural Language Processing." NLP-progress, 2022. http://nlpprogress.com/. (NLP-progress 2022) This describes the 'state-of-the-art' in the subfields of natural language processing.

Chapter 10: What are the Opportunities for Librarians?

10.1 Introduction

In 1989, Edward Feigenbaum ('the father of expert systems') observed in the paper *Toward the library of the future* that the problem with the then extant libraries was that the books did not talk to each other (Feigenbaum 1989). He continued:

> ... imagine the library as an active intelligent knowledge server. It stores knowledge of the disciplines in complex knowledge structures, perhaps in a knowledge representation formalism yet to be discovered or invented. It can reason with this knowledge to satisfy the needs of its users. These needs are expressed naturally with fluid discourse. The system can, of course, retrieve and exhibit. That is, it can act as an electronic textbook, but it can also collect relevant information, it can summarize, it can pursue relationships. It acts as a consultant on specific problems, offering advice on particular solutions, justifying those solutions with citations, or with a fabric of general reasoning. If the user can suggest a solution or an hypothesis, it can check it. It can even suggest extensions, or it can criticise the user's viewpoint with a detailed rationale of its agreement or disagreement. It pursues relational paths of associations, to suggest to the user previously unseen connections. Collaborating with the user, it uses its processes of association and analogizing to brainstorm for remote or novel concepts. With more autonomy, but with some guidance from the user, it uses criteria of 'interestingness' to discover new concepts, new methods, new theories, new measurements (Feigenbaum 1989, 122).

Feigenbaum is addressing a certain kind of library here, what would be called an 'academic library' or a 'research library' (and those categories might include university libraries and medical libraries).

There are many types of libraries, including:

Academic libraries
Children's libraries
Digital libraries
Medical libraries
National libraries
Public lending libraries
Reference libraries
Research libraries
Special libraries
University libraries

[See (Wikipedia 2023i; American Library Association 2007) for a description of some of these and an explanation of their functions.] The librarianship activities associated with these are many and varied, and there are also other librarianship activities not connected with institutions specifically of these types. For our purposes, as a practicality, we have to restrict our gaze. For the most part, we will be looking at machine learning in connection with scholarship, research, and advancing knowledge (i.e. with Feigenbaum's approach of relating AI to the notion of a library as a knowledge server). This means that our main focus will be academic libraries (including university libraries), medical libraries, and research libraries. We also often consider librarianship in general. There may be the odd remark on machine learning in other kinds of libraries (as examples, that humanoid robots may be valuable for children's story times in public libraries (Nguyen 2020), and that handwriting recognition may be valuable

for the Vatican Archive, which is a special library (D. Firmani et al. 2018)), but coverage of these areas is going to be thin. Sorry.

There is the idea of collections as data and data as collections, with librarianship as an interface (see, for example, (Padilla et al. 2019)). Standard libraries can be thought of as collections— collections of texts, documents, and books, and also, perhaps, of means of access to the same or similar items. Some, or many, of these collections, or parts of collections, will be born digital, or become digitized, and will be available to computers, artificial intelligence, and ML. Thus, there is the notion of collections as computer data. But also, we now live in the age of big data. Huge amounts of data are being accumulated by researchers, governments, social agencies, and commercial interests. Many subsets of this big data are collections. They are libraries. They are subject to the ordinary concerns of librarianship, such as: organization, preservation, storage, access, retrieval, and stewardship. So, there are collections as data and data as collections, with librarianship as an interface. For example, researchers in astrophysics, with their telescopes, radio dishes, and myriad of other instruments, produce data repositories that stand in need of librarianship. Then, still in their day jobs, these researchers read research papers in journals and collections. Provision of these also needs librarianship.

There is an abundance of modern digitized, or born-digital, resources, and this abundance is growing rapidly all the time. There is a plethora of sources spewing ever more 0s and 1s. Facing up to this on behalf of librarianship are a relatively small number of expert human librarians. There is an order of magnitude difference here— the potentially valuable

tasks, and the collection sizes, far outweigh the capabilities of a team of human librarians, even if the number of librarians were to be increased a million-fold. As an example, which is becoming slightly dated, there is the MeSH indexing of biomedical publications (about 7000 articles a day were being indexed). Yuqing Mao and Zhiyong Lu wrote:

> MeSH indexing is the task of assigning relevant MeSH terms based on a manual reading of scholarly publications by human indexers. The task is highly important for improving literature retrieval and many other scientific investigations in biomedical research. Unfortunately, given its manual nature, the process of MeSH indexing is both time-consuming (new articles are not immediately indexed until 2 or 3 months later) and costly (approximately ten dollars per article) (Mao and Lu 2017).

There is a general point to be made here (the same general point as will be made elsewhere over and over). Computers have a 24x7 work ethic, and ML can often supply expertise. Many areas have been automated already. But this is an ongoing and expanding process. Of course, human librarians use many tools to increase their capabilities viz-a-viz librarianship, but we are moving into an age where the ML systems on their own can produce excellent performance (maybe even superior performance).

Librarians already work with many systems that have connections to AI and ML. Here is one way to classify librarians working in AI or ML in roles:-

Librarians can be seen as

- Synergists
- Sentries

- Educators
- Managers
- Astronauts

Individual librarians may fulfil different roles on different occasions, or, indeed, be working in different roles at the same time.

10.2 Librarians as Synergists

Librarians have several thousand years of experience of working with recorded information. They are ideal partners to AI to bring out the best in this information with all its aspects, challenges, and facets. And, on the other side of this, there are many AI and ML technologies that have the potential to improve librarianship. There is an opportunity for synergy or symbiosis.

We will introduce a few sample possibilities here and expand on the topic later in a dedicated chapter.

Synergists:

- **Intellectual Freedom.** AI, ML, and librarianship have potential to enhance Intellectual Freedom, both the expression of free speech and access to it: as examples, Optical Character Recognition leading into machine reading and speaking of text, and, separately, machine translation of text leading to recorded text being available in many

languages (Knox 2023). Machine learning is now expanding these possibilities to hundreds, even thousands, of languages.

- **Smartphones.** Many young, and not so young, people use smartphones as their main means of access to information. This opens an obvious invitation to librarians to bring libraries and librarianship to patrons via smartphones (e.g. via a bundle of voice and sound, static and dynamic text, and images and video). This might involve AI, ML and chatbots.
- **Improving Intermediation Between Users and Information Resources.** As examples: search engines rank their returns, from more important to less important, and, separately, there are 'recommender' systems which can recommend other resources similar to ones favored by the User. Pure librarianship does not have either ranking or recommending in any developed form. Even the computer, or AI systems, versions have not yet realized their potential.
- **Improving Traditional Cataloging, Classification, and Retrieval Tools.** Standard point: there are, and always will be, fewer expert human catalogers than are required to address the increasing flood of resources. Computers and ML can redress this. Also, NLP can perform valuable librarianship tasks that are not practical for humans. For example, it could look at a million publications and identify, *de novo*, what subjects, topics, or genres each might be labeled with. 'De novo' here means 'without using any antecedent classification schemes'.
- **Chatbots.** Chatbots have the potential to do most of the current tasks where librarians interact with patrons, either synchronously or

asynchronously (e.g., in person, on the telephone, on video calling, by messaging, etc.).

- **Release, Produce, Curate, or Inspire the Production of, Training Data.** Librarians already have metadata on the contents of libraries— metadata which, for the most part, is accurate and labeled well. Also, librarians are well placed to use handwriting recognition on archives of historical documents. Developing handwriting recognition for these will require the production of data (most likely labeled samples).
- **Social Epistemology.** The promotion of social knowledge— social epistemology— is a vital function of librarianship (Egan and Shera 1952; Fallis 2002; 2006; Fuller 1988; Goldman 1999). It is a function that librarians have been doing for millennia. Social epistemology faces problems aplenty nowadays with disinformation, misinformation, fake news, deep fakes and the like (Meszaros and Goodsett 2022). Librarians, in conjunction with the tools of ML, are well placed to take on the challenges. There is now the opportunity, and the need, to do more.
- **Images.** There are libraries and collections of images (for example, the Center for Creative Photography in the University of Arizona (CCP 2020)). Also, many publications contain images, figures, and diagrams. Attaching metadata to the images, and then finding the desired images, has always proved difficult. But now ML is allowing it to become possible for standard librarianship operations to be performed on media like images. For example, you can now enter an image itself, or a verbal description of an image, as search 'terms', and a suitable search system will be able to find all similar or relevant

images within a document or collection. These possibilities would be enriched by insight from librarians. Images are getting to be addressed reasonably well by ML.

Jason Griffey asks an interesting question in connection with synergy. He introduces it by way of recommender systems and improving personal intermediation:

> … the system trains itself from the user's behavior. One can easily imagine systems built to do this sort of automation work for researchers and students. As AI systems continue to be easier to implement, having a system local to your device that learns your preferences, your interests, and your needs will be commonplace. Researchers and students will have AI systems that find sources for them, summarize them, help them build bibliographies, and more. Over time, these systems will become irreplaceable archives of the learning and thinking history of individuals, a sort of universal diary of their activities. Now, imagine for a moment that this sort of system exists and is used by most learners. *Who would you prefer be the developer of such a system: a large corporation like Facebook, or a collaborative effort by educational institutions and libraries?* (Griffey 2019, 27) [Italics added.]

We will see an example later of 'a collaborative effort by educational institutions and libraries'— that by Kent Fitch in the Section 10.2.4 (Fitch 2023). Another example should perhaps be mentioned, that of Kaushik Roy and fellow authors using retrieval-augmented generation (RAG) (Roy et al. 2023)

10.3 Librarians as Sentries

Unfortunately, many of the potential benefits of ML and librarianship have concomitant downsides. Here the librarians can be sentries. To anticipate, the challenge is that advances in ML have been so rapid that suitable ethical systems, laws, and policies either do not exist or are out of date. Librarians can help create these.

Sentries, here are some examples:

- **Copyright and intellectual property** Intellectual Freedom interacts with restrictions of privacy, intellectual property, state secrets, and so forth. These considerations required careful management (and librarians have plenty of experience with, for example, intellectual property, fair use, and licensing).
- **Bias management** There are various kinds of bias that can arise in connection with ML and information provision. Librarians do have experience in managing bias. For example, they do it in collection development, collection description, instruction, and research support (Padilla 2019).
- **Monitoring techniques to improve search** Methods associated with personalization and recommendation impinge on privacy (by, for example, monitoring Users' behavior to create the personalization). Filtering has had a bad reputation within librarianship (primarily due to misadventures involving filtering schools' access to websites). But, when doing a search, providing

recommendations is filtering. Then filtering can lead to information silos or bubbles. There are problems here to be addressed.

- **Intellectual freedom** This needs management. The collections and services should presumably give patrons access to a wide range of diverse and thought-provoking materials, while also protecting them from potentially harmful or offensive or false or ungrounded or obviously crazy content. But how to do this is a question: one person's crazy might be another's happy territory. And the idea of 'protecting' introduces paternalism which should not be required for fully functional adults. Machine learning systems could very easily work in a paternalist way.
- **Inadvertent censorship** The properties and behaviors of advanced machine learning systems for example, those built from foundation models, are usually not fully known. Caution is needed to ensure, for example, that there is not accidental censorship.

10.4 Librarians as Educators

Librarians have a role as educators.

Educators:

- **Information Literacy** Librarians have always been the standard bearers for information literacy. But Artificial Intelligence (AI) has changed what information literacy can be. There is ongoing development of new tools for interacting with information, for example, personalized search. AI or ML, as research disciplines or

commercial enterprises, devote little or no attention to information literacy itself (except in so far as the AI or ML can be a part of any educational course or teaching on any discipline or topic).

- **Data Literacy, Data Science Fluency, and AI Literacy** There are other forms of information related literacy that are becoming important (Ridley and Pawlick-Potts 2021a; Digital2030 2022; Druga et al. 2019a; Carlson and Johnston 2015; Padilla 2019). For example, research scientists are often required to have data management plans. They are producers of data, and they need to know how to manage it for the benefit of other researchers and the world at large. Librarians can help the researchers directly and also play a role in educating student researchers in the management of data. Another example is that AI and ML have expanded the realm of Automated Decision Making (ADM) (e.g. the making mortgage loans). An informed citizenry should be alert to the strengths and weakness of ADM.
- **More Intelligent Consumers of Information**. This includes both patrons and library staff.
- **Better Informed Citizens** Outside of actual information literacy, there are considerations of helping citizens understand ML and computational aspects of the world they live in (and, in the case of living in the USA, how some other countries are approaching it). As examples, there is the Canadian *Algorithm and Data Literacy Project* (Digital2030 2022), and there is the European *Generalized Data Protection Regulation* (Wolford 2018).

10.5 Librarians as Managers

At a perhaps a more day-to-day level, librarians run libraries, both physical and digital. Computer assisted automation is widely used and is of obvious benefit. Book acquisitions, cataloging, serials control, and circulation, information retrieval and dissemination, interlibrary loan, cooperative acquisition and cataloging have been automated in the library (Lakshmikant and Vishnu, 2008).

AI can improve the running of libraries. We are trying to steer clear of plain automation in this text. We will try to restrict ourselves to cases where the software uses or simulate artificial intelligence.

Managers:

- **Workflow and Improving Service** ML has the potential to enhance productivity and efficiency in libraries. Many the components here have been mentioned already: ML cataloging, personalization, recommender systems, better search, chatbots for customer service, predictive analysis for collection management, user behavior analysis to improve service, and digitizing special collections.
- **Optimize the Use of Space (and, Indeed, Other Resources)** ML is good at optimization problems.
- **Robots** To put books back on the shelves (!), to do story-telling, to meet and greet, and more.

- **Mimic Librarian Experts' Behaviors** To support decision making and management.

10.6 Librarians as Astronauts

Astronauts, well, who knows? But most of human knowledge is in libraries. ML will allow exploration here of a kind that has never been done before.

Astronauts:

- **Creating Knowledge**. There is deep text extraction and synthesis from materials already in libraries. More than a few university researchers conduct their research using only their initiative and the contents of libraries. ML will be able to do this (and render the faculty researcher redundant in this regard).
- **Drawing Out Knowledge**. There are many special collections that have not been digitized and transcribed (and, perhaps, for some of those that approach might not be acceptable). But processed collections— with indexes, for example— might provide access to treasures.
- **Moonshots ? Who Knows What They Might Be?**

10.7 Annotated Readings for Chapter 10

[Several of these publications are out of date, as are many sections of the present text.]

Asemi, Asefeh, Andrea Ko, and Mohsen Nowkarizi. “Intelligent Libraries: A Review on Expert Systems, Artificial Intelligence, and Robot.” Library Hi Tech 39, no. 2 (2020): 412–34. https://doi.org/10.1108/LHT-02-2020-0038. (Asemi, Ko, and Nowkarizi 2020) This is a reasonable literature review. (Some of their references seem incorrect as to their topics e.g. it identifies Amin and Razmi 2009 as being on the topic of 'knowledge-based indexing'.)

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

Cordell, Ryan. “Machine Learning + Libraries.” LC Labs. Library of Congress, 2020. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf. (Cordell 2020). This has a good survey of topics, and a valuable bibliography (generously also provided as a Zotero shared library). There seems to be no mention or discussion in it of chatbots.

Cox, Andrew M., and Suvodeep Mazumdar. “Defining Artificial Intelligence for Librarians.” Journal of Librarianship and Information Science, 2022, 09610006221142029. https://doi.org/10.1177/09610006221142029. (Cox and Mazumdar 2022) This brings a different conceptualization to the interactions between AI and libraries to the one offered here.

Cox, Andrew M., Stephen Pinfield, and Sophie Rutter. “The Intelligent Library: Thought Leaders’ Views on the Likely Impact of Artificial Intelligence on Academic Libraries.” Library Hi Tech 37, no. 3 (2019): 418–35. https://doi.org/10.1108/LHT-08-2018-0105. (Cox, Pinfield, and Rutter 2019). There is a useful table in this of possible AI initiatives, relevant competencies, and 'alternative providers'. As to the latter— commercial interests, publishers, or university Information Technology departments may produce or provide the AI tools or services. Librarians watch out!

Dempsey, Lorcan. “Generative AI, Scholarly and Cultural Language Models, and the Return of Content.” LorcanDempsey.net, 2023. https://www.lorcandempsey.net/generative-ai-a-note-about-content/. (Dempsey 2023b)

Das, Rajesh Kumar, and Mohammad Sharif Ul Islam. “Application of Artificial Intelligence and Machine Learning in Libraries: A Systematic Review.”

ArXiv:2112.04573 [Cs], 2021. http://arxiv.org/abs/2112.04573. (R. K. Das and Islam 2021).

Fernandez, Peter. "'Through the Looking Glass: Envisioning New Library Technologies' How Artificial Intelligence Will Impact Libraries." Library Hi Tech News 33, no. 5 (2016): 5–8. https://doi.org/10.1108/LHTN-05-2016-0024. (Fernandez 2016). Library Hi Tech News has a column, occasional articles, written by Peter Fernandez. These are good, recommended.

IFLA. "IFLA Statement on Libraries and Artificial Intelligence," 2020. https://repository.ifla.org/handle/123456789/1646. (IFLA 2020) This is particularly good, recommended. It has useful references in its Annexures.

Jakeway, Eileen, Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, Abigail Potter, and Kate Zwaard. "Machine Learning + Libraries Summit Event Summary," 2020. (Jakeway et al. 2020). There seems to be no mention or discussion of chatbots in this.

Padilla, Thomas. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." Dublin, OH: OCLC Research, 2019. (Padilla 2019). This has input from around a hundred knowledgeable practitioners and academics. It aims to develop '… a research agenda to help chart library community engagement with data science, machine learning, and artificial intelligence'. As such it is slightly different in aspiration to the present text (which tries to look at the intellectual challenges arising between ML and librarianship). It appears not to contain a mention or discussion of chatbots.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. "Always Already Computational: Collections as Data: Final Report," 2019. https://doi.org/10.5281/zenodo.3152935. (Padilla et al. 2019)

Rolan, Gregory, Glen Humphries, Lisa Jeffrey, Evanthia Samaras, Tatiana Antsoupova, and Katharine Stuart. "More Human than Human? Artificial Intelligence in the Archive." *Archives and Manuscripts* 47, no. 2 (2019): 179–203. https://doi.org/10.1080/01576895.2018.1502088. (Rolan et al. 2019) This is directed at archives and record keeping.

Chapter 11: Librarians as Synergists

11.1 Intellectual Freedom

Intellectual freedom is the right, and perhaps the ability, to access or disseminate any information that the person wishes. Now, there are many qualifications here, obviously, such as those concerning privacy, intellectual property, offensive materials, state secrets, etc. Let us build in the restrictions and consider only those cases where there is a right to freedom of access or to freedom of speech (i.e. to freedom of dissemination). Then let us concentrate on the *ability* to exercise our rights in those cases. It is reasonably common in rights discussions to distinguish *privilege* rights from *claim* rights (Wenar 2021). A person has a privilege right to do X, if, and only if, they do *not* have a duty *not* to do X. So, a mature person presumably has a privilege right to read, say, Huckleberry Finn, and that right exists because they do *not* have a duty *not* to read it. Additionally, a privilege right is a freedom, to the right's holder, which does not impose any obligations on any other party to facilitate that right. If you have a privilege right of freedom of access to some form of information, you may access that information, but no one is obliged to give you practical help to do so. Claim rights, in contrast, impose a duty on some person or entity to actively assist a person to exercise the right. For example, some folk might not be able to afford to buy, or simply might not wish to own, some of the books that they are interested to read— in these cases, libraries, perhaps public libraries, can facilitate a claim right of access to those resources. In contrast, book sellers or bookstores are not subject to a claim right of access

(in the absence of purchase, or similar, from the would-be reader)— they do not have a duty to provide access.

Libraries, especially public libraries, aim to satisfy claim rights of access. Some also may provide a physical public space for meetings. This would be assisting with disseminations, perhaps even assisting with claim rights concerning dissemination. Libraries, or librarians, also often are active in opposing 'banned books'. Third parties might want copies of, say, Huckleberry Finn, removed from a library. What is happening here is that the third parties are partially denying the claim rights of others (and possibly also denying the privilege rights of others). Librarians are defending these rights. So, librarians help meet some intellectual freedom rights and defend others (Garnar and Magi 2021; American Library Association 2008).

How machine ML help with this? The short answer is: librarians need to provide direction as to what they wish their policies to be by way of addressing rights of access and dissemination. In the past they have provided such direction, and they continue to do so in the present. But the landscape as to what is possible is changing rapidly, so policies require ongoing attention. ML cannot really help with what the policies should be. But once the policies are in place, ML can certainly help with the implementation. Here are some details.

11.1.1 Text Recognition

Text recognition has been an absolute triumph in enabling people with visual challenges, or reading difficulties, to access written and printed materials. In a library setting, text is understood to be recordings that are written, or printed, or reproduced, pieces of language that can convey meaning. These recordings can range from handwritten documents to professionally printed books, from documents thousands of years old to those from the present day, from originals to photocopies of photos of photos pasted into a painting collage. The recordings have a form, and mostly they have meaning. Text recognition is trying to capture the structured form of the text as computer input, and not the deeper meaning. For example, consider the text ‘Now, I am angry.’ What text recognition needs to be able to do is to pick the letters in sequence i.e. ‘N’, ‘o’, ‘w’, etc., to pick the words (likely from the separators such as the blank spaces), to pick the sentences (likely from the upper-case letters, periods, and punctuation), etc. It does not need to be able to grasp what the text means (e.g. what or who the indexicals like ‘I’ and ‘now’ refer to, or what the word 'angry' means).

Two further points should perhaps be made. Text recognition's main line of business is not that of helping people who are visually challenged, rather it is that of helping large corporations and institutions convert their paper records and data into digital form. Assisting the visually challenged is happy side effect of that enterprise. Second, complexities arise with OCR over whether the source documents are monolingual or multilingual, and

consequently the various alphabets, orthographies, and grammars that might be in use (see, for example, (Alpert-Abrams 2016)). It is possible for target sources to be multilingual. They would constitute additional challenges. Multilingual sources are not common. But they certainly can appear in older materials where there might be a mix of a scholarly language (e.g. Latin) with a vernacular (e.g. Italian, French, English.) Consider an older handwritten text, written in several languages, and which uses 'loan words' (i.e. uses or quotes words from other languages). Adequate OCR here might require relevant outside knowledge (for example, of the provenance, and cultural background, of the document).

As mentioned earlier, most modern publications exist at some point in digital form. If there is access to that digitization, text recognition is not needed. So, generally speaking, text recognition for library purposes is not as important now, 2023, for modern materials, as it was, say, for materials produced prior to the 1980s. Also, the use of a variety of fonts and printing styles, which make text recognition harder, are a modern development. That help is to some degree offset with other problems with older materials such as difficulties with ink fading, poor printing, deterioration of the paper, etc. Nevertheless, the main problem areas for text recognition in libraries are older printed works, where the publishers have a restricted choice of font palette, and, separately, handwritten documents. Generally, printers are trying to print works in a form that the readers can read. They are not trying to distort letters left and right, or larger and smaller in way that would challenger readers (or, indeed, OCR systems). The original OCR systems of the 1970s would not have been ML systems, but nowadays they

certainly would be. For a discussion of OCR, we can consider just neural nets.

Conceptually, text recognition takes place in two phases: the optical character recognition (OCR), and the discernment of structure within the stream of characters. From an ML point of view, OCR is an example of a supervised classification problem. First there is a classification system, which is a collection of classes, or categories, or sets— in this case, that is going to be an alphabet of characters e.g. ‘a’, ‘b’, ‘c’ etc. Then there will be the characters, or character instances, themselves which will start computational life as visual disturbances, marks, blobs, etc. on paper or some other medium. Likely they will then become patterns of pixels in electronic images that the algorithms can address. Processing will individuate these into characters-of-interest, then ML will classify each one of these as being an ‘a’ or a ‘z’ or a ‘j’ or a space or other character.

It sounds simple, but it is not. There are issues and challenges. Let us start with the classification system. There are about a dozen writing systems, and hundreds of alphabets (Ager 2023). There is the need to pick one or several here. Which one is chosen presumably depends on the purposes at hand. If a library has primarily English sources, presumably the English Alphabet would be a good start. This is not in any way to insult or denigrate, for example, Japanese and Hiragana. There are two points to be made here. Classification systems usually carry baggage with them. They have assumptions and consequences that extend further than the classification systems themselves. (See, for example, (Bowker and Star 2000; S. Berman 2000).) An OCR implementation might be able to recognize English

characters but not Hiragana— and that might matter for those who can or cannot benefit from the specific OCR. Second, the choice of classification system or systems might involve a wide range of constituents, not just programmers or AI researchers imposing their will. The patron, user, librarian, research challenge, and infrastructure and legal framework, also can provide input.

Then Supervised ML OCR is going to be taught, or learn, how to classify characters. It will be supplied with a *training set*, which will be a reasonable sample of letters and the right labels or classifications of what they are. The training set might run to 100,000 labeled characters. The overall technique is an optical one, so it is the features of the sample letters that can be detected optically that will be the input (e.g. size, shape, color, grid arrangement of component dots or pixels, etc.). Then the program will attempt to correlate combinations (i.e. vectors) of these with the correct classification e.g. that a particular sample token character is an 'a'. More than likely, the program will make many mistakes initially. But either the programmers, or the program itself, will tune various parameters (e.g. weights on the components of the vectors) to improve the classification until it reaches an acceptable level of performance. Typically here, the neural net would have 5-6 layers and hundreds to thousands of neurons in the layers.

The training set needs to be adequate for the task. For example, if the letter 'j' does not appear in the training set, it is unreasonable to expect the ML program to classify js correctly. Even if js appear, there needs to be enough of them in the various fonts and scripts (cursive or not, monospaced or

proportional, etc.) for the program to be able to learn what is correct and what is not. OCR, i.e. the task of recognizing the actual individual characters, would not usually be an end in itself. Rather, the interest would be in the words that those characters form, or, more generally, the text.

If the OCR, or Text Recognition, application has access to a wider context, that can improve its performance. For example, if the ML is recognizing entire words from their component characters, and separators, then the first letter of 'On' is going to be the letter upper case 'O' and not the numeric letter zero '0'— the number zero makes no sense in that context.

Let us assume going forward that there is ML that can take an input of images (i.e. a page of visual representations of characters, an ordinary book of words, etc.) and produce as output text. Now, text, in a computer science sense, consists of 'strings' or sequences of characters, and, once in that form, they can be processed in a variety of ways. For example, they can be searched, or edited (cut, copied, pasted, transformed etc.). To give a practical example of the advantage of strings over raw images, finding the word 'covid' in some (computer science) digital text is near trivial for a computer program. Contrast that with the following. Imagine a photocopy, or photograph, of the front page of a newspaper, and the problem finding whether there is a sub-image in it that might be construed as an image of the word 'covid' — i.e. does the word 'covid' appear on the front page of the newspaper? That is a much harder problem. (You know that, of course, from the online Captcha tests that are used to detect whether you are a human or a program or 'bot' pretending to be a human (Wikipedia 2023a).)

OCR for modern printed monolingual text is near perfect. There is a qualification here. OCR needs training data. There are about 7,000 languages in the world, and about half of these have writing systems. Of those 3000 or so with writing systems, quite a lot less have enough printed text to be suitable training data. Current OCR systems can read about 200 different languages. There will be more languages than that with suitable training texts, but there also needs to be either commercial or intellectual incentives for the relevant OCR research to be done.

As the Text Recognition systems have improved, their compass has been extended to include handwriting recognition i.e. transcribing handwriting to the 0s and 1s of computer text. This is very important. To give an example. The Vatican Apostolic Archives (the Vatican Secret Archives) contain hundreds of thousands of documents going back many centuries (D. Firmani et al. 2018; Wikipedia 2022f). Most of these documents are handwritten, and certainly more than a few of them are of great significance. As examples, one is from Henry VIII to the Pope requesting a marriage annulment, there is the Catholic Church's 1521 excommunication of Martin Luther, and there are notes from the trial of Galileo.

Transcribing handwriting is of a level of difficulty harder than transcribing printed text. There are different cases to be considered. There is personal handwriting to be transcribed 'online' (i.e. as it is being written, real-time, perhaps onto a smartphone or tablet). For example, a User may handwrite entry into a text-messaging app. There is purely personal, or official, handwriting to be transcribed 'offline' (i.e. from a recorded document after it has been written). Transcribing online is easier than offline because

information is available on the pen strokes, their sequence and timings. This information helps, for example, with segmentation: with identifying the lines, the words, and the characters. There can be real-time transcription with the characteristics of offline, for example, so-called scene-in-scene transcription. For example, driverless cars may have the need to 'read' road signs and other textual information that they are 'seeing'. There may also be some value in the ability of a driverless car to read handwriting within its video stream. Some restaurants display a menu outside with the day's fare handwritten on a small blackboard. Search applications might want to have the ability to read this from a video feed. For our purposes, with recorded documents, our interest is primarily with offline transcription.

> The problem of automatic Handwritten Text Recognition (HTR) persists since document digitization started. Text recognition is a simple task for humans, but it has been proved to be complex for automatic systems. In fact, it is considered an unsolved problem and under active research.... The high variability between writers and the cursive nature of the handwriting text are the main difficulties presented by this problem. These difficulties have meant that historically, the practical applications of offline handwriting recognition technologies have been quite limited (Sueiras 2021).

To a degree, everyone's personal handwriting is different. Nevertheless, mostly, people are trying to communicate with their handwriting, and they have learned, have been taught, or are required, to write in a way that their writing can be read. There are differences in the writers, their writing styles, and their purposes. There are differences in the intended roles of the product documents.

Some cases are easier to transcribe than others. There is cursive handwriting and block handwriting (writing separate letters). For example, international arrivals at an airport may be required to hand 'print' (i.e. block handwrite) their flight and passport details into a form, where some of the fields of the form are required to be text, others known to be dates, and yet others known to be numbers. Such handwritten forms can be machine read quite easily, even though there may be different authors and writing styles involved.

Documents like those in the Vatican are often official documents of one kind or another, written by scribes. In cases like these, not only does the scribe personally want the document to be easy to read and definite in form, but so does the scribe's employer (the government, the Queen, the seller of the land, etc.). Scribes need to get it right or lose their jobs (or maybe their heads). Scribes in a certain cultural setting are usually required to follow a specific style (roughly: they have to write in a certain font). Most of the medieval and later documents in the Vatican will have been written in the Carolinguan Miniscule font (or later fonts related to it).

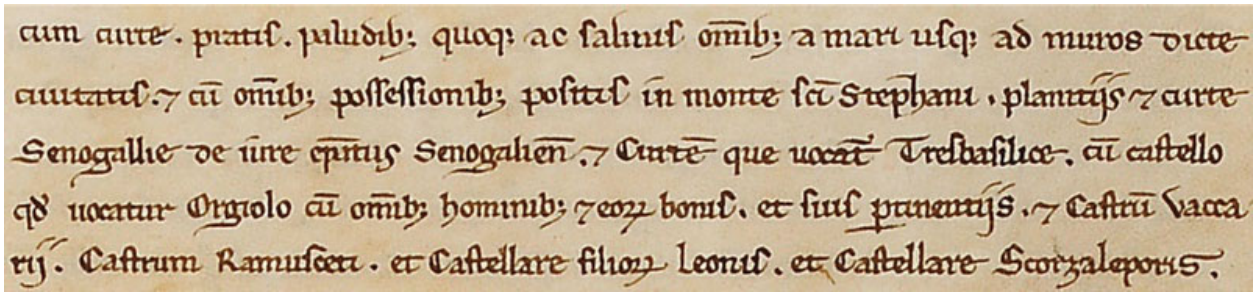

Figure 25. Sample Text From the Manuscript "Liber septimus regestorum domini Honorii pope III", in the Vatican Registers (Donatella Firmani, Merialdo, and Maiorino 2017).

Many cases are much more difficult than this. There is a manuscript collection of the works of the English philosopher Jeremy Bentham. (Bentham wrote on philosophy and law. He is known mainly for proposing Utilitarianism.) Bentham wrote the manuscripts largely himself. But he also had helpers who wrote some portions. He wrote mostly in English, but occasionally with pieces in Greek or French or other languages. He made corrections, but often leaving the originals and the crossings out in the text. He also sometimes wrote in columns, or with included passages.

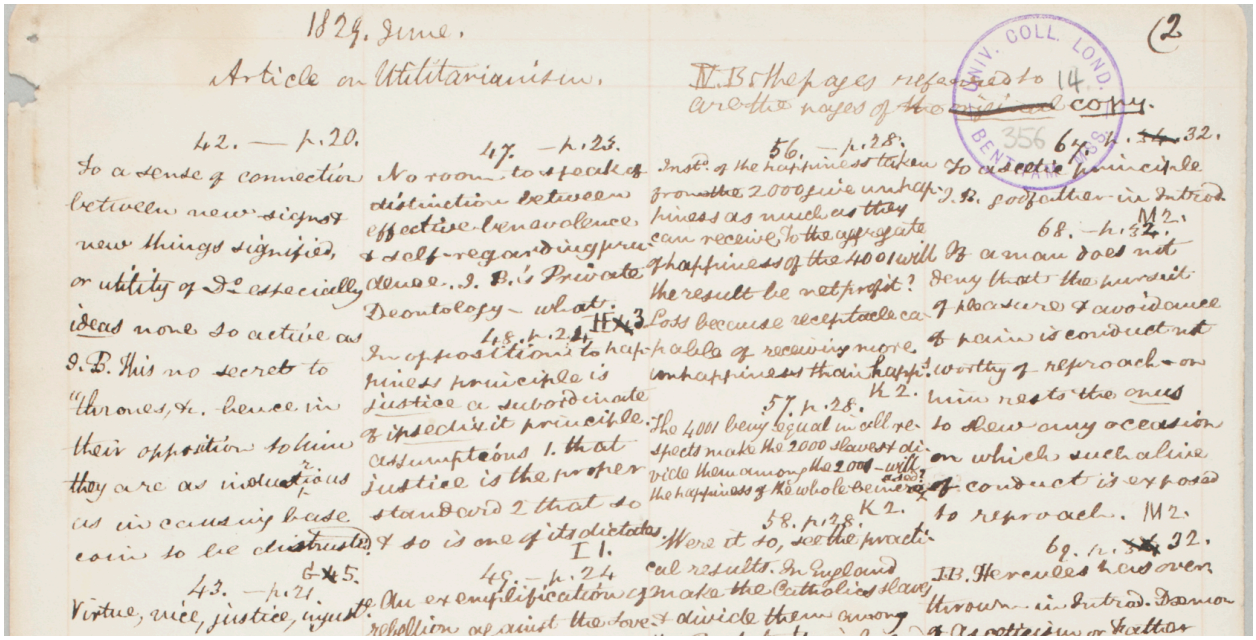

Figure 26. Sample From a Manuscript by Jeremy Bentham (University College London).

However, coming at this from a different direction, think back a few years of your own experiences. A regular High School teacher might receive thirty handwritten essays from students and be able to read them without any difficulty. (If there was difficulty, likely the student would be given remedial handwriting instruction.) This suggests that ML would allow computers to read cursive or block handwriting from a variety of authors. Typical approaches to this in 2023 would likely involve Neural Nets or Deep Learning, and this means, basically, that they could be covered by a Foundation Model.

11.1.2 Speech to Text

We are all familiar with voice-controlled assistants like Alexa, Siri, or Google Assistant. These can take speech, or dictation, and first turn that into (computer) text. Usually, the way that this step is done is that an ML program will attempt to recognize and classify the phonemes in the input sound. The phonemes are the smallest, or atomic, sound components of speech in a language. In English there are about 44 of these. In response to dictation, the program will produce a stream of phonemes. Converting this stream into well-formed and sensible text is hard. But it has been done now with 99% accuracy for most well-known common languages. To give some idea of the difficulties. In English, there can be the same letters producing different phonemes (‘thin’, ‘these’), different letters producing the same phonemes (‘sit’, ‘city’, ‘eight’, ‘ate’). To resolve these, context is required—i.e. what makes sense in the wider sample of speech. Google's Cloud Speech-to-Text V1 supports dictation in hundreds of languages (Google Cloud 2023). Apple’s software for iPhone covers dictation capabilities for about a hundred languages or their dialects (British, American and Australian English, are all listed separately, for example). Meta's Massively Multilingual Speech (MMS) project is aiming at speech recognition and text-to-speech models that can recognize over 4000 languages and work with over 1,100 languages (Meta 2023). Meta also open-sources its models (which may be good). Of course, recognizing a language is one thing, being able to take dictation in it at speed is something else.

The training here is largely with audio recordings which have transcripts (obviously here resources are meagre with rare or unusual languages). If the target language has a text-to-speech converter, that can be used to create data.

Speech-to-text can be used to transcribe speeches and lectures. Text-to-speech can be used to create audiobooks. The two technologies can be used to help people learn languages. The two technologies can be used to help people communicate in their preferred language. They can help preserve language diversity and prevent languages from 'dying'. Languages usually die from a lack of use and native speakers. This may come about for a variety of reasons (speakers favoring other languages, government policy, etc.). However, the technologies create an infrastructure which retains the knowledge of how a language is to be spoken, written, and read.

There are settings where either typing or writing cannot be used or in which they are not the best option. Dictation is usually taken to be much faster than typing or writing, maybe 3-10 times faster. [There is evidence counter to this in certain circumstances such as start to finish form-filling in a medical or legal context.] Some situations— for example, if there is background noise— are not the best for speech input. But the choice is there. Voice commands, voice input, and the ability to ask questions verbally increase intellectual freedom.

Smartphones are ubiquitous. They can take text input and provide visual text output (and some users are astonishing adept at typing their text

input). They also can work with sound and will usually have a voice-controlled assistant.

Transcribing spoken language into text makes it easier to search. This may be useful for libraries that have audio resources (such as historical recordings of Question-and-Answer sessions to a parliament, or oral histories.)

11.1.3 Sign Language to Text, and Text to Sign Language

There are AI programs to translate sign language to text. These will use some means of capturing or 'viewing' the images, perhaps a smart phone. Then they will classify what is seen. This may be done real time— that is, as the same speed as the signs are being given i.e. at conversation speed. But classification of, say, the hand movements is not the only challenge. There are maybe 300 sign languages, and they have different characteristics, of course. Usually, a sign language does not consist solely of signs made with the hands. Rather there is a wider range of gestures including facial expressions and bodily movements. This makes the translation problem much more difficult.

Within libraries, and other repositories of information, text, images, and other visual artifacts are the primary display medium. If a person is challenged visually, then text-to-speech technologies are very helpful. If a person is challenged aurally, but not visually, then sign-to-text technologies are certainly valuable. But the visual text itself would presumably be

available to such a person, and that would provide a means of access to recorded information. There will be cases where sign-to-text technologies would help in a library setting— for example, for those who can sign but not read or write. Libraries do have holdings in sound; for example, podcasts, recordings of lectures and speeches. Transcription to text might help there.

We are all familiar with important addresses and speeches being signed simultaneously with being spoken. Obviously, that is a valuable access and inclusion technique. But when the focus moves to recordings, say of the Australian Prime Minister, Julia Gillard, saying, in 2012, 'I will not be lectured about sexism and misogyny by this man...' , there does not seem to be the same need for a signed version. Recordings of the speech are available in video, sound, and text.

11.1.4 Helping Filter and Personalize

A problem highlighted by Jorge Luis Borges's short story *The Library of Babel* is that just having access to absolutely all information would likely prove to be useless because the relevant information would not be able to be found (Wikipedia 2023o). Intellectual freedom needs a filtering of garbage, and a personalization to provide relevance. ML can provide that. [This topic will be revisited later.]

11.1.5 Scholarly Publishing

This is a small aside, but it is an example. Researchers in the western world are more-or-less required to publish their research papers in English. Many are not native English speakers. There is an awkwardness at this point. Some scholarly publishers provide automatic translation to English for all submissions (see, for example, (CHOICE Media Channel 2022)).

11.1.6 What Can Be Done With Computer Text

Suppose, now, that we have the computer text corresponding to a page, or pages, or books of printed or written characters, or of speech, or of signs. What can be done with that? A lot, see Section 1.4.

11.1.7 ELI5 Translation

It is common to see the acronym 'ELI5' on discussion hosts like Reddit. ELI5 means 'explain it like I am a 5 year old'. An interesting and fascinating point is that LLMs can do this. They can explain passages as though the audience were 5-year-olds. But also they can simplify and re-render passages at any level whatsoever from 5-year-olds to the level that the passage was written at original (and even go in advance of that perhaps to a more sophisticated version).

This matters for two reasons. It increases intellectual freedom, because it makes difficult passages accessible to all, and it personalizes the delivery of information into a form that the user would like or to a learning style that the user needs.

11.2 Improving the Intermediation Between 'Users' and 'Information Resources'.

11.2.1 Some Users Might Not Be Human

Some Users might be ML programs, or software tools that are employed by the human User, or employed by other programs or 'bots'. To a degree, this is true already. For example, many repositories will publish their metadata for 'harvesting'. (See, for instance, the *Open Archives Initiative Protocol for Metadata Harvesting* (Wikipedia 2023k)). What is happening here is that services are being built off resources that a library or libraries have, or off services that libraries offer. Librarians can help this by ensuring that their holding are accessible in relevant ways to appropriate 'Users'. Accessibility in this context will likely involve considerations of licensing and intellectual property. One type of service that likely will become very important is that of text mining. We will look at text mining later.

11.2.2 Some Resources Might Not Be Resources

Some resources might not be resources in the sense of being physical or digital texts or documents. They might be services. A library might provide a service, for example a certain kind of access to its holdings. A second library might provide a similar or different services. A further party might bundle those services into another service.

11.2.3 Digital Archiving

Archiving is a little different to straightforward librarianship in that it deals almost exclusively with historical materials and its organizational principles are often different (for example, in giving pride of place to provenance). Digital archiving— preserving digital content for future use— is different yet again. It is one of the many things that commercial and governmental institutions do. For example, digital archiving may be mandated for compliance reasons— e.g. keep your tax records for 10 years!

11.2.4 Enhanced Search Engines

Existing search engines already use machine learning. Companies like Google typically do not reveal exactly how their systems work— presumably for commercial reasons. But some techniques are known or can be surmised. Google search uses RankBrain, Hummingbird, Panda, Penguin,

Pigeon, etc., along with the original PageRank algorithm (Wikipedia 2023l; 2023m; 2023f). Some, or maybe most, of these use ML. What these techniques concern mostly is with ranking the results that the search engines find. Apparently, there are 200 or so ranking factors that can be taken into account. Some of these are to do with the web pages themselves (whether they contain real content or are 'click bait'). Some concern the users and their search histories. Some even concern the locations of the search origins. For example, in the US the word 'boot' picks out a footwear item you wear, in Britain the word 'boot' can mean this but it can also mean what Americans would call the 'trunk' of a car. The search engines know this and can react accordingly. The search engines also use NLP more generally, and there is considerable amount of machine learning in NLP.

Commercial search offerings are close to being able to do the following in response to a query:

1. Semantic search, augmented by named entity recognition, to produce and return, in the first instance, say, ten links to relevant web pages
2. To summarize those pages at a length and intellectual level either suitable for, or requested by, the user
3. To load the contents of those pages into a LLM ‘database’, thus allowing the user to chat and ask questions about the web page contents, with the LLM answering (with citations)
4. To construct a knowledge graph from the contents, and from this produce topics or references of further interest. It could here produce a list of related searches.

The technologies in use here are:

- Semantic search (from NLP, embeddings, and LLMs)
- Summarization (from NLP and LLMs)
- Question answering (from NLP, probably vector stores of embeddings, and LLMs)
- Knowledge graphs (these are infrastructure in this setting— what they are is explained in Appendix E)

One example of a 'librarianship paper' showing work of this kind is Kent Fitch *Searching for Meaning Rather Than Keywords and Returning Answers Rather Than Links* (Fitch 2023). Fitch writes:

> Large language models (LLMs) have transformed the largest web search engines: for over ten years, public expectations of being able to search on meaning rather than just keywords have become increasingly realised. Expectations are now moving further: from a search query generating a list of "ten blue links" to producing an answer to a question, complete with citations.
> This article describes a proof-of-concept that applies the latest search technology to library collections by implementing a semantic search across a collection of 45,000 newspaper articles from the National Library of Australia's Trove repository, and using OpenAI's ChatGPT4 API to generate answers to questions on that collection that include source article citations. It also describes some techniques used to scale semantic search to a collection of 220 million articles (Fitch 2023).

Librarians have expertise in information retrieval. This could be used to create search engines that are more effective at understanding and fulfilling user queries. There already are image searches, with image input or image

output. There are location-based searches. Google maps tells you of restaurants near you, not those on the other side of the continent. There are possibilities and opportunities here.

11.2.5 Personalization and Recommendation

What an information provision system is trying to do in response to a perhaps not well articulated request from an individual User is to supply that User with all and only the relevant resources (i.e. 100% recall and 100% precision) from a well-defined, or not so well-defined, collection of items. [Recall is the percentage of relevant items in the collection that are returned. Precision is the percentage of returned items that are relevant. Roughly, recall is signal and precision is absence of noise.]

There are explanations and qualifications required here. Martin Frické (2013) writes:

> Relevance is usefulness to the User as judged by the User. Relevance is just a user-controlled honorific that connects [items] and utility (on a particular occasion of retrieval). The Patron or User is ultimately the sole arbiter of relevance.

Assume so. This means relevance is subjective in the sense of being User and occasion specific. In turn, this means that retrieval is subjective, and the best retrieval systems will allow for personalization for individual Users. Separate from this, the boundaries of a collection may be somewhat wooly as they may range from a single bookshelf out to a single library from there out to the entire Internet.

ML can personalize provision and retrieval by knowing about a patron and a patron's past behavior. Traditionally, libraries have often had advisory services. These usually would consist of one or more librarians who would know the local collection and would have experience of patrons, their information needs, and the resources from the local collection that would meet those needs. The assessment or feedback on their work might be limited. There might be some surveys, or similar, but not much more.

Computer and AI 'advisory services' would or could know about many users, many collections, and would have extensive data and feedback as to how it was performing (and improve itself accordingly).

11.2.6 Recommender Systems

We are familiar with recommender systems from our experiences with Amazon, for books or other items, and with Netflix and streaming sites for movies. Recommender systems can improve reading experiences.

There are various possibilities here as to how these might work, which we can describe in a library setting. Data is needed, and pretty much the more the better. This might include:

- demographics about the User (e.g. whether they are a child or a senior citizen)

- any information that they wish to share about their likes and dislikes and preferences (genres, subject matters)
- their reading or access history
- information about the books or available resources (perhaps including their genres, abstracts, ratings, or reviews).
- other relevant factors
- diversity in plots, characters, authors, genres

There is collaborative filtering which puts the user in the context of other similar users and recommends on that basis. There is content-based filtering which pays attention only to the items that the user likes, or seems to like, and works with their properties (ignoring information about other users). Most systems will use a hybrid of both approaches. One is that an anonymous user accesses or views or reads a single item and remains anonymous (but might have a continuing single session identity). There is not much data to work on here. But the item itself will have many properties, such as author, subject, genre, length etc. It also might tie into explicit written reviews or feedback from other patrons or even professional reviewers or critics. Some sorts of recommendations might be able to be made here. The continuing session identity might give some feedback as to whether any of the recommendations were followed up on, by the central user, during that session.

More usual would be the setting where every patron's accessing history, and every item's accessed history, is known and recorded as data. (This data can be kept private, and identities not revealed.) This accessing history will be 'implicit' data about the items accessed. There also will be 'explicit' data

about these items, such as their authors, genres, subject matters, reviews, and citations or references or links to them. At this point, either just data about past interests might be used or data about that supplemented by data about other users and their histories. Any ongoing behavior by a user can be used to update their profile.

Then, probably, either one of two approaches might be made. The first is to put the patron into a 'stereotype' (i.e. a class or group) consisting of other patrons similar in respect to the patron seeking recommendations. This would be done largely on the basis of the present reading and the access history. The other is not to bother with classes and just to let a ML system look for similarities in reading behavior across users. It takes the likes and dislikes of the User being helped then overlays those on the likes and dislikes of other individual Users to produce a match. The upshot can be a recommendation system that helps users with personalized, perhaps ranked, recommendations as to resources that would be interest, useful, and relevant.

Such systems can do more. They might be able to predict how new items, yet to be purchased items, will be received by groups of users— the preferences. And thus, in the case of libraries, they can help with collection development.

11.2.7 Understanding What the User is Asking For

Internally, behind the curtain so-to-speak, a traditional information retrieval system will likely use Boolean queries or queries in a database query language like SQL. But few Users are competent to do input their questions or requirements in this fashion (see, for example, (Frické 2021)). Some Natural Language Processing here could smooth the interface between User and such systems.

11.2.8 Text Mining

As described in Chapter 9, text mining can extract a variety of potentially useful information from text such as information regarding named entities, topics, classification features, genres, summaries, and sentiments of reviews. Any or all of these may help a User narrow an information search and thus make it more precise.

Going a little broader, text mining can look through (usually large) corpora for valuable information or patterns. The size of the task makes it hard, if not impossible, for humans to do. We have discussed facets of this elsewhere— for example, creating an encyclopedia by ML requires text mining. Question answering, summarization, tracking research ideas, etc. all require text mining. We will discuss the topic again in the context of Undiscovered Public Knowledge.

One red flag or alert is over the question of licenses or the legal position over the mining of texts. In so far as they can, libraries ought to ensure that they can mine their holdings. (Unfortunately, the reality might be that some other entity will do the mining and charge the libraries for doing so.)

11.2.9 Information Assistants (and 'GPTs')

Let us adopt a form of thinking here. Let us characterize information intermediation in terms of *tasks* and *control* (or *flow*). Tasks consist of searching, recommending, paraphrasing, translating, etc. Control (or flow) is how the tasks fit together. There is *sequential* flow, which is where tasks follow each other in a sequence. There is *conditional* flow, which is where there is a condition (call it 'if') and if the condition is satisfied (i.e. is true) flow goes down one branch (one further sequence of tasks) and if not the flow goes down another branch. Finally, there is *loop* flow, which is where a sequence of tasks repeats or loops either a given number of times or until a condition is satisfied.

Given this structure, informal information algorithms can be constructed. For example,

> Is there a day of the week that the Musée d'Orsay in Paris is closed in July?
> Is there a day of the week that the Louvre in Paris is closed in July?
> If they are both closed on the same day, say what day that is, otherwise say which museum is open when the other is closed and what the relevant days are.

> Look up recent research on twisted spin.
> Summarize the best papers, no more than 5 papers.
> Present the summary at the level understandable by a graduate student in physics.

We can conceive of these in terms of flow and tasks, and so can LLMs. LLMs can take this kind of input in English, spoken or written, and answer it.

[Editorial Note. The first edition of this book continues:

> "The answers, July 2023, may be a bit rough. But it will be only months before the answers will be very good.
> In sum. Shortly there will be information assistants that can combine information tools on the spot. Users will be able to mix and match tools. That might not matter to library patrons on all occasions of their uses of libraries. But the lives of researchers are going to be transformed."]

On November 6th 2023, OpenAI announced 'GPTs' and the upcoming GPTs Store which will sell GPTs or provide them free. As mentioned earlier in Section 2.9, there is a builder technology that allows the construction of GPTs. GPTs themselves are relatively small assistants or agents based on the underlying LMM GPT-4 Turbo technology (or its successors). Assistants work as partners with humans. Agents are autonomous and once given a task or project do not need further human input before completion. Present GPTs should probably be classified as assistants, but agents are only the blink of an eye away.

11.3 Improving Traditional Cataloging, Classification, and Retrieval Tools

The elephant in the room here is presented succinctly by Tamar Sadeh in her doctoral thesis and a series of papers including 'From Search to Discovery' (Sadeh 2015). The argument is: the traditional approach required users to learn library systems and articulate the perfect template to launch a search which would then be guaranteed to produce a perfect result straight off. In contrast, the modern user could not care less about library systems. In their daily lives, they use Google search and do online shopping all the time. They enter the information discovery process in a sloppy and haphazard way. But get some, or many, results which are then honed to meet their needs. The process is familiar to them. Sadeh describes this:

> The designers of traditional library information systems, such as library catalogs and databases, were very focused on meeting the needs of librarians and expected that users would invest time and effort in learning how to use the system. The designers of discovery systems, driven by the needs of end users, strive to streamline the end-to-end process of finding and obtaining information and make it as simple and friendly as possible. Rather than offering multiple options to enable users to describe their information need, discovery systems offer users simple search interfaces but complement these with multiple post-search options for assessing findings, refining results, and navigating to other results of possible interest. The look and feel of the interface is similar to that of other information systems that are familiar to users, such as web search engines and online bookstores. Furthermore, recognizing that today's users spend hardly any time reading instructions, developers have made discovery systems very intuitive. (Sadeh 2015, 216)

So, we can look at the topic improving retrieval tools, but some improvements may be for librarians only.

There is the view from Patrick Wilson and Karen Coyle that traditional cataloging theory omits the User from its concerns (Coyle 2016; Wilson 1968; Svenonius 1969). This view invokes *descriptive power*, describing the resource items that libraries have, and *exploitive power* which evaluates and recommends items suitable for patrons or Users on particular occasions. Cataloging does the former, but not the latter to any competent and enthusiastic degree. There is subject classification, but that has not been carried out very well (S. Berman 1971; Frické 2012). There are other library services and tools that help with exploitive power, for example: bibliographies, reference interviews, and similar. But librarianship to date has been weak on exploitive power. Relatively new computer supported search engines, with ranked returns, and 'recommender' systems are strong in these areas. But ML, especially large language models, have the potential to take this to new levels.

Thomas Padilla writes

> ...semantic metadata can be generated from video materials using computer vision; text material description can be enhanced via genre determination or full-text summarization using machine learning; audio material description can be enhanced using speech-to-text transcription; and previously unseen links can be created between research data assets that hold the potential to support unanticipated research questions (Padilla 2019, 12)

This is exactly right. We will supplement this in places and give detail to some of the suggestions.

Over the millennia that librarianship has been practiced, librarians have developed many retrieval tools. Here are a few, supplemented with some modern techniques:

Abstracts
Bibliographies
Book reviews
Catalogs
Citation Indexes
Computer Interfaces
Controlled Vocabularies
Cumulative Indexes
Databases
Dictionaries
Encyclopedias
Finding Aids
Handbooks, manuals, etc.
Indexes
Inventories
Keywords
Nomenclature
Ontologies
Outlines, syllabi, etc
Pathfinders
Reference Interviews
Reference Lists
Registers

Reviews
Search Engines
Single Entry Term, Phrase, or Keyword, Search Boxes
Subject Guides
Summaries
Tables of Contents
Textbooks
Thesauri
Web Browsers

As we will see, ML will likely be able to improve most of these tools individually (see, as examples, (Iris.ai 2023; Pickering et al. 2022)). A different consideration is whether ML can replace some of the tools entirely.

11.3.1 NLP Inspired Improvements

Most of the AI/NLP areas of value to librarianship are mentioned and described in Chapter 9:

- Named Entity Recognition
- Topic Modeling, Text Classification and Automatic 'Tagging'
- Controlled Vocabularies, Thesauri, and Ontological Vocabularies
- Automatic Indexing
- Text Abstracts, Extracts, Key Phrases, Keywords, and Summaries
- Sentiment Analysis

- Author and Genre Recognition and Plagiarism Detection
- Text Mining and Question Answering
- Machine Translation

In general terms, these all make improvements to older techniques. But some bring features that are genuinely new:

- Topic Modeling can identify new topics in huge corpuses of texts e.g. in social media.
- Indexing can identify sources outside of the original indexed document. This could be useful for research.
- Sentiment Analysis can be useful for recommender systems. For individual books, for authors, for genres. For individuals, for the patrons as a whole ('trending books') and for collection development.
- Author and Genre Recognition and Plagiarism Detection.
- Text Mining and Question Answering.
- Machine Translation.

11.3.2 Metadata Generation and Automatic Cataloging

Machine learning can certainly be used to create all the forms of metadata that are in use in librarianship today i.e. it can do cataloging. (See for example (Griffey 2019; Corrado 2021)). What would be likely here is interactive reinforcement learning or human-in-the-loop learning. That is, during the training process professional catalogers would provide feedback

as to how well the automatic system is performing. The catalogers would be part of the training. (Of course, it is not easy to know what good cataloging amounts to (Snow 2017).)

In certain circumstances, automatic metadata generation has the potential to be very useful. For example, UNESCO has audio recordings that would benefit from having metadata. They have about 6500 recordings, in 70 languages, and some of the recordings are multilingual. What they have done in the past is that an intern listens to a recording and picks topics and personalities. There are ML systems that can do the speech recognition and transcription (for example, Whisper from Open AI (OpenAI 2022b)). Apparently one of the challenges these systems can have is with crosstalk (e.g. meetings where several different people talk as and when they feel so inclined).

11.3.3 Some Retrieval Tools

The opportunities or possibilities here are extensive. We will address just a few examples.

Producing a *List of References* that are actually cited in a text is a trivial computer science problem. There are many human-powered citation managers (e.g. EndNote, Zotero, and Microsoft Word), and all of these can produce 'Bibliographies' (here meaning reference lists, or citation lists). True *Bibliographies* are another matter. A 'True Bibliography' is being

understood in this context as being a list of the works used to write or produce the text, whether *those works are cited or not*.

Producing a true *Bibliography* is no easy task. The author or authors of a work can do it. But for third party humans or computer programs, it is a challenge. Computational linguistics can classify texts, can identity the genre, can identify whether two works were written by the same author, can identify plagiarism, etc. But these methods, whether they use ML or not, mostly rely on what is in the text explicitly. Locating what is in the background that might have inspired the explicit text is a challenge of a different level entirely. As of 2023, it cannot be done.

Paula Carina de Araújo and fellow authors and, separately, Linda Smith, provide a good introduction to the vast area of *Citation Analysis* and *Citation Indexes* (Araújo, Castanha, and Hjørland 2021; L. C. Smith 1981). In brief, a reference is an explicit acknowledgement that one text gives to another, and a citation is reference received by a text from another. Citations and references can be used to build webs or networks between documents. Such networks, which often amount to citation indexes, can be valuable for a variety of scholarly purposes. Centrally

> … (1) they are tools for the scholars seeking knowledge and (2) they are tools for the scholars studying science (including scientometricians and information scientists). (Araújo, Castanha, and Hjørland 2021)

One point to note about traditional citation analysis is that what is being considered is actual citations made by human authors. These citations may

be made for a variety of reasons. (As Lizzie Gadd notes, the *Citation Typing Ontology* lists 43 different types of citation (Gadd 2020; Peroni and Shotton 2012).) Let us pick a semi-random five of these: *is confirmed by*, *corrects*, *critiques*, *derides,* and *disagrees with*. Consider a particular research paper, A, and another research paper B. If the human author of A is conscientious and knowledgeable, she may cite B for any of the five, or other, reasons. But, nevertheless, given the vagaries of life, she may not cite B at all. However, the paper A may still be confirmed by B (or correct B or critique B etc.). That is what the paper may do, objectively. A scholar of ideas, of intellectual history, of the development of the theories in a field, may mainly be interested in the relationships between A and B, not in whether the author of A cited B. Additionally, consider the time before it was the practice to make citations. Ancient Greek or Roman authors wrote texts that, for example, critiqued other texts. Human creators of citation indexes or analyses basically can only work with actual citations. But at this point ML and NLP systems have a crucial advantage. They can scan the entire research literature and form a knowledge or information map of which papers confirm which other papers (or deride which other papers, etc.). They can scan the content (as well as the gossip of actual citations). New knowledge mapping tools are, and will be, far superior to their traditional counterparts (see, for example, (Tay 2022)).

Moving on. ML systems could write *Book Reviews*. With fiction, it could identify plots, characters, themes, whether the content was 'diverse', intended audience, and other aspects of the document or book. With non-fiction, it could assess quality by means of coherence, 'groundedness', truth-and-evidence, writing style, citations it uses, citations to it, and other

indicators. Also, it could, using sentiment analysis and other NLP techniques, collectively assess, summarize, and evaluate reviews written by other agents (human or otherwise).

Libraries make extensive use of *Databases*. Almost all information about their own individual holdings will be held in databases. Access to these will often be via their *Catalog*, which might be in the form of an *Online Public Access Catalog*. Also, there are any number of commercial and other databases that serve as access points to further resources outside an individual library's holdings (for example, to research papers, to legal materials such as citations and precedents). An academic library might provide access to hundreds of outside databases. Of course, databases will be used in the everyday administration and management of libraries (such as for patron and circulation records, for staff salaries, etc.).

Database theory is a specialist area, widely studied in computer science. Databases themselves provide organization to their contents. They also should be able to do so in a structured and provably correct way. They support CRUD operations (Create, Read, Update, and Delete). There is, or can be, plenty of automation in connections with databases, for example, with checking the data on entry (for format, reliability, etc.), checking integrity, following a backup or archiving or compliance policy. It is not clear quite what role machine learning might have in this domain. ML can program a database. It can do any computer programming approaching the level of professional programmers. How good it would be a design is an open question. This is an area where there are many formal techniques— Entity-Relationship diagrams and the like— and ML would presumably

easily master those. Machine learning can have plenty of relevance to the contents of databases, picking up patterns in the data for one reason or another (for example, identifying fraudulent transactions in a financial database). Somewhat similarly in a library setting machine learning will be able to identify usage patterns in the resources— for example, which resources are used by which segments of the patrons— and thus aid acquisitions and collection management.

Traditional *Catalogs* list all the materials held in a library. The lists will have data and metadata about the resources. If a library physically issues books and materials, a catalog might assist with lending, checking availability, placing holds, etc. All of these functions benefit, or will have benefitted, from computer automation. The old timey favorite *Card Catalog* was a technology to help the patrons find items among those materials listed in the Catalog. As computers, networks, and automation came in the Card Catalogs evolved into *Online Public Access Catalogs* (OPACs). OPACs steadily acquired additional functions such as access to materials in other libraries or in other formats like databases. OPACs are probably drifting off into the sunset (Wells 2021). There is a better alternative, the so-called *Discovery Systems* (such as Primo VE, Summon, or EBSCO Discovery Service). Tamar Sadeh writes:

> Discovery systems provide access to a large, diverse information landscape of scholarly materials regardless of where the materials are located, what format they are in, and whether the library owns them or subscribes to them. At the same time, these systems typically offer simple, Google-like searching as the default option, to accommodate the expectations of today's users. With this type of searching, users do not spend much time formulating queries, and their queries often yield large result sets; therefore, discovery

> systems focus on relevance ranking and on tools that help users easily navigate and refine result sets. Librarians have welcomed the advances in discovery services for their users. However, this new reality poses challenges to the practices that librarians have developed over the years and, in some cases, is at odds with the systematic, controlled approach to searching endorsed by librarians (Sadeh 2015).

(with, for example, personalization and recommendation— such as Primo VE, Summon, or EBSCO Discovery Service).

LLMs can create *Dictionaries* — after all, they will have seen massive amounts of text in its natural contexts. Many dictionaries provide examples in use. LLMs would be able to provide richer and more comprehensive examples. Right now, though, 2023, it would be usual to construct dictionaries by editing existing dictionaries. Merriam-Webster's dictionary, for example, is about 200 years old, and about 1000 new words are added each year. Words are removed also. There are corpora— collections of real world text— for example, the Open American National Corpus (anc 2023). Computer analysis of corpora tells which words are new and how frequently they appear, and also which words are drifting out of use. Then, presently, human judgement, assisted by computers, makes decisions on how to edit the dictionary. The whole process could surely be done by ML and LLMs on their own. The LLMs in question would presumably have some downstream training from human experts. The commercial companies— such as Merriam-Webster— do use artificial intelligence, as examples to personalize the experience to the User and to provide usage examples and notes.

There are *Encyclopedias* that have in part, or in whole, been created by ML

- *Wikipedia* uses AI … 'This [use] may be directly involved with creation of text content, or in support roles related to evaluating article quality, adding metadata, or generating images' (Wikipedia 2023p)
- Wikimedia is using ML to help with images for *Wikidata* (Redi 2018)
- *Encylopedia.com* (Encyclopedia.com 2019*)*. This is an access point, rather than an encyclopedia in itself. It gives access to 200 other encyclopedias and can search and summarize.
- Numina Group's *Warehousing 'Encyclopedia'* (NuminaGroup 2023) (This is more of a glossary or catalog.)

ML created encyclopedias could or should be completely up to date, accurate, and comprehensive. They might be expensive, biased, and with some entries that were hallucinations.

A *Pathfinder* is:

> A subject bibliography designed to lead the user through the process of researching a specific topic, or any topic in a given field or discipline, usually in a systematic, step-by-step way, making use of the best finding tools the library has to offer (Reitz 2014).

The finding tools mentioned here include: catalogs, bibliographies, indexes, abstracts, bibliographical databases, and search (by author, title, topic, and keywords).

Machine learning can improve pathfinders in many ways:

- The whole task of producing a pathfinder can be done automatically.
- It can personalize a pathfinder to a User, instead of their being a single pathfinder for many Users.
- Its search will be better.
- It can find and follow topics even when those topics do not have their own metadata by having been catalogued by the library systems.
- Indexes will be better.
- Abstracts will be better.

11.4 Chatbots

Chatbots have been in use in libraries for at least ten years (McNeal and Newyear 2013; Weigert 2020). They are improving. They have the potential to do, or interface to, most of the current tasks where librarians interact with patrons. Most obviously here are Reference Interviews. There may be a loss of some personal touch. But the gain might be a tireless expert reference librarian for every patron, available 24 hours a day 7 days a week. Another possibility is that of supplementing Frequently Asked Questions (FAQs). After all, at the start of a freshman year at a college, there may be several thousand students wanting to ask the same routine questions (such as the opening hours for the campus libraries). Chatbots might also produce a thinning or abandonment of present-day library websites. A deeply structured website will often not be the best way of directing users to the resources on offer (Wikipedia 2022e; Thoppilan et al. 2022).

There may be a loss of personal touch. There may be a gain of personal touch— some folk are really enchanted with chatting with chatbots. There is some evidence that chatbots outside a library setting do not provide a good 'customer experience' (ujet.cx 2022b; 2022a)(Standard point: one benefit might be indefinitely many tireless expert reference librarians, enough for one for every patron, available 24 hours a day 7 days a week.)

Librarians would be valuable, perhaps even necessary, in the creation of suitable chatbots. Likely the chatbots will use the GUS architecture with frames with slots (Bobrow et al. 1977). (The frames provide the contexts, and the slots the values of the variable for the data such as the questions and answers.) Librarians have a better idea than most what the frames should be for a librarianship setting.

11.4.1 Reference Interviews

Some librarians either are reference librarians, or, as part of their duties, conduct reference interviews. Librarians here are acting as intermediaries between patrons or Users and reference or information sources. A few years ago, the aim of a reference interview was to match a patron's needs to a single library's resources. Nowadays the resources would be assumed to extend outside a single library perhaps to all libraries and, indeed, to the Internet itself.

Several sources provide the same instructional guide to librarians as to how to conduct a successful reference interview. Here is a typical outline:

Purpose:

Allows staff to match the customer's question to a relevant and useful source of information. The aim of the interview is to answer the patron's questions using the library's resources.

Guide To A Successful Reference Interview:

1. Approachability
2. Interest
3. Listening/Inquiring
4. Searching
5. Follow-up

1. Approachability

Pay attention to both your own and the customer's body language. Acknowledge and greet the customer as they approach the desk. Ensure the customer has your full attention.

2. Interest

Maintain eye contact
Find a confidential location for the customer to ask a question
Restate and rephrase the question
Speak in a relaxed tone
Make the customer feel comfortable
Nod your head when the customer starts to ask questions

3. Listening/Inquiring

Do not interrupt
Ask clarifying questions
Let the customer express their needs in their own words
Ask open-ended questions to probe about their information needs
Examples:

Tell me more about the sources of information you already consulted?
Why do you need the information?
How will you use the information?

Remember, WORF

Welcoming, Listen Carefully
Open-Ended Questions

Repeat their answer back to the customer
Follow-up to ensure they've found the information

4. Searching
    Keep the customer informed of the progress
    Offer referrals
    Offer to instruct the customer on how
    Ask clarifying questions:
        Do you want printed information that you can take home with you?
        Do you have access to a computer so you can look up sources online?
5. Follow-up
    Asking the customer if they have everything they need ensures that the customer is satisfied with the transaction.
    If the follow-up questions indicate that the customer is not satisfied:
        Clarify what information is missing
        Offer to continue working on answering the question
        Refer the customer to another organization if material is not available at your library

[The author dislikes the use of the word 'customer' but maybe that is just him.]

ML and LLMs can certainly excel at all this. Joseph Vincze has a useful discussion in his paper 'Virtual Reference Librarians (Chatbots)' (Vincze 2017). [A caution: that paper was written before the advent of ChatGPT.]

11.4.2 Virtual Services

There have been Virtual Reference Desks, Ask-a-Librarian web pages, and Chat-with-a-librarian through a web page, more-or-less since the internet expanded and became popular. Some of these have been retired, for

example the Library of Congress Virtual Reference Desk was retired around 2020.

The LLMs complete change the game on this. You can have ChatGPT on your smartphone. That is a virtual reference desk, without needing a library or librarians. As of June 2023, ChatGPT is not perfect with accuracy, providing references, and avoiding hallucinations. But it is improving all the time, and there are good reasons to suppose that ChatGPT with appropriate plugins will outperform any extant virtual reference desks. Librarians might work with ML researchers to create the plugins.

11.4.3 Chatbots as Continuous User Testing of a Library's Public Interface.

This might be controversial, and it certainly would need handling carefully. But chatbot transcripts would throw good light on what patrons actually do, or plan to do, while using a library's resources. For example, a web page with a large number of show/hide toggles (accordion widgets) may prove hard to follow; a chatbot leading a user through the page could provide feedback on this.

11.5 Release, Produce, or Curate Training Data

ML is only as good as its data. Training data is hard to come by. It needs to be plentiful, and it needs to be of high quality. For supervised learning, the

data needs to be labeled accurately. Librarians are well placed here with all kinds of suitable data. They already have rich metadata on traditional resources, including shelf classification, subject classification, and indexes. And they can produce new kinds of data. For example, the READ-COOP projects to use handwriting recognition on archives of historical documents (READ-COOP 2021). These projects can involve creating and inspiring crowdsourcing to produce data at scale, e.g.

> Transcribe Bentham is an award-winning participatory initiative which launched in 2010. Its aim is to engage the public in the online transcription of original and unstudied manuscripts written by Jeremy Bentham, his correspondents, and his amanuenses. (UCL 2018)

(For an explanation of crowdsourcing see (Wikipedia 2023c).)

Public librarians can inspire crowdsourcing. Many modern ML projects, especially Large Language Models, and Foundation Models, use crowdsourcing in their training. For this they generally need 'ordinary' people, but collectively the crowds would usually need to be diverse (as to race, ethnicity, religious beliefs, sexual orientation, etc.). Public libraries interface with many hundreds of thousands of exactly the kinds of folk that would be suitable, possibly even ideal.

11.6 Debunking, Disinformation, Misinformation, and Fakes

Librarians have always been active with information literacy, which is helping users and patrons to become more discerning and skillful in their approaches to information resources. (We will discuss this again later.) But, in addition to this more general raising of skills on the part of users there is or can be a scrutiny of the actual resources themselves. There are problems aplenty nowadays with disinformation, misinformation, fake news, deep fakes and the like (Meszaros and Goodsett 2022). Traditionally, librarianship would not pass a view on the veracity of sources or on evidence. Librarians would remain neutral, for example, between resources on evolutionism and on creationism (and there are good free speech arguments for doing this). But there is considerable value now in fact checking.

Librarians already help patrons to fact check e.g. (Knapp 2021). Many libraries and library associations are deeply involved in fact checking Really this is a part of epistemology, perhaps social epistemology.

11.7 Social Epistemology

Margaret Egan and Jesse Shera introduced social epistemology as being:

> ... the analysis of the production, distribution, and utilization of intellectual products in much the same fashion as that in which the production, distribution, and utilization of material products have long been investigated.(Egan and Shera 1952)

One core part of this concerns what we know— the true beliefs we have—as individuals and collectively. Most of what we know individually comes from other people via recorded knowledge. Certain practices regarding that recorded information can promote, or inhibit, social knowledge (i.e. knowledge aggregated across individuals). As an obvious example, easy wide access to recorded information promotes social knowledge whereas censorship inhibits it. Some qualifications are needed to this example. Too much information might overwhelm our attention and interest. A little censorship, or filtering, or curating, might highlight the pearls among the dark sea of many biased and unsupported opinions. Egan and Shera, and other later researchers, such as Don Fallis, Steve Fuller, and Alvin Goldman, have seen the promotion of social knowledge as being a vital function of librarianship (Fallis 2006; 2002; Fuller 1988; Goldman 1999; Egan and Shera 1952).

Librarians are experts at traditional information acquisition, and information provision practices. That is a good start. But machine learning is both going to provide more powerful tools to help with social knowledge and, perhaps a mixed blessing here, a vast amount more source material for those tools to be used on. Here is a conjecture about recorded text, especially new materials on the Internet and on Social Media. The Large Language Models, such as ChatGPT, Bard, and Bing, can write English, and other languages, as well as native speakers. They can do so quickly, much quicker than native writers, and cheaply, much cheaper than native writers. Very shortly, tools from these models will be in the hands of anyone that

wants them (as tools on the web or in word processors or stand-alone apps on smartphones, tablets, or computers). What the tools will produce will be plausible in terms of vocabulary and grammar. Some of the textual products will be information. Others will be misinformation. Some will be 'hallucinations'. Some will be fiction, intending to be fiction. Some will be fiction, not intending to be fiction. It seems that ML source creators will be as ubiquitous as spell checkers are today— every means of producing content will have available an LLM assistant.

What might be roles for librarians in connection with ML and social epistemology? Here are some possibilities:

- Fact checking. Help with identifying misinformation, disinformation and 'false facts'.
- Help with cognitive biases. Many folk have trouble with reasoning— with hypotheses, evidence and truth. Indeed, one experiment showed that 50% of Harvard Physicians can commit the base-rate fallacy. (The base-rate fallacy is explained in Appendix C.2.) AI can help keep people on track. AI can construct proof trees from arguments and evidence. There are other relevant cognitive biases. For example, the phenomenon of confirmation bias suggests that almost everyone mistakenly favors evidence that supports their views, downplaying or ignoring disconfirming evidence. ML and LLMs do not cause this, but they can be used to counter-act it. For example, for patrons interested in balanced view of, say, climate change, librarians, using LLMs could find the confirming and disconfirming evidence. The librarian’s role here would be part synergist, part sentry, and part educator.

- There is the notion of Veritism, or truth-centered epistemology (Kitcher 2002; Nawar 2021), and from this the question arises of whether social epistemology needs to be veritistic. There is the need for input from philosophy. Once there is an acceptable answer to this, ML may be able to help.
- There is the view, now becoming widespread, that Peer Review as used, for example, in scholarly publishing has failed. One proponent of this is Adam Mastroianni (Econtalk 2023). ML can do everything that peer review is supposed to do: check spelling, grammar, citations, diagrams and figures, calculations, originality, contribution to the research field, absence of plagiarism, etc.

11.8 Robots

For convenience here, we will divide robots into three categories: chatbots, humanoid robots, and non-humanoid robots. [There may be some overlap of these boundaries; for example, a humanoid robot might have chatbot capabilities.] Then, orthogonal to this, a physical library may be *using* robots, or providing *access to* robots.

Chatbots are discussed elsewhere (and there are many opportunities for librarians with chatbots). Humanoid robots are beginning to be introduced as companions and helpers to the elderly in rest homes. There are also some uses in nursing. Such robots are often mobile and can converse. The 'humanoid' part usually includes humanlike expressions of emotions and gestures. These can engender trust and reduce anxiety on the part of the

people the robots are interacting with. Somewhat similarly to being helpers in rest home, robots have been trialed in public libraries. As examples of actual uses, being a teller of stories in story-telling sessions for children, being a greeter to the library and answering directional or locational questions to books or resources (Nguyen 2020; Kim 2017). Non-humanoid robots, for example, welding robots used in the manufacture of cars, would usually appear as part of the automation of processes. There has been inventory control in general, using RFID (Radio Frequency Identification) labels and chips, since the 1980s. This means that it is relatively easy to automate closed stack systems (where the public do not browse the books on the shelves). The physical collection can be 'in the basement' and automation will do the rest. RFID, and similar technologies, are also invaluable with open stack systems. A hand-held scanner, or a wand, or even built in systems, in the shelves or walls of the building, will find, for example, any book or identify that it is missing. Increasing use of automation may be useful, but there does not seem to be a large role for robots. As to libraries providing access to robots, robots are going to be an important part of our future. This suggests that librarians can help educate the populace by, for example, lending robots or having makerspaces with access to robots or having educational seminars on robots.

In sum. It is unclear quite what might happen in general with humanoid and non-humanoid robots in libraries, and what the opportunities might be. (See also (Tella 2020; Tella and Ajani 2022).)