Humanities and Human-Centered Machine Learning

Abstract

Machine learning research focuses extensively on evaluation, but much less on user intentions and values. We argue that there are two primary paradigms of machine learning use: one that builds independent agents, and another that builds interactive tools. When researchers do not recognize differences in intentions and their associated value systems, it can lead to conflicts and misunderstandings. In this chapter we highlight case studies drawn from computational humanities research to illustrate the characteristics of these two interaction patterns.

1 Introduction

The language of machine learning (ML) research is saturated with evaluation — metrics, benchmarks, and baselines — but strangely devoid of user intention. What is ML for, and how will it be used? In this chapter we argue that there are two distinct paradigms of user intention: tool-building and agent-building; that these paradigms have distinct value systems; and that within the research community one of these paradigms is so dominant as to make the other one functionally invisible.

Much of ML research is focused on an *agent-building* paradigm, in which a system is intended to replicate the activity of a human. The goal in this paradigm is to eliminate human interaction, such that when the system is presented with a set of inputs, it will produce the same outputs that a human would have produced given those same inputs. For example, a classifier may be used to automate a decision process. Its goal is to take a new example and replicate the label that a human annotator would have applied to that example. The implicit assumption is that the correct answer is known or at least knowable, and a good system will reduce incorrect or uncertain predictions. One reason this paradigm is attractive to ML researchers is that evaluation is simple: you create a collection of actual human input-output pairs, and measure how well your system produces the specified "gold standard" or "ground truth" output. Once the system matches human performance, you can presumably insert the system into whatever process the human was previously doing (or not doing). Consideration of human interfaces is thus minimized or invisible.

Outside of ML research, the goal of ML applications is often not to erase human interaction but to enable it by creating a *tool*. The goal is insight, rather than operational capacity. This tool-building paradigm centers iterative, expert human use: we are creating an instrument that will be actively used by an expert with substantial domain knowledge to carry out complex tasks that support a research goal. Model output may be incomplete, inconsistent, or even wrong, and yet still be of value because it will be used by an expert user who is able to evaluate and contextualize it. This pattern is more difficult to evaluate with automated metrics, as it centers complex human interactions.

Use paradigms and the humanities In this chapter we use the work patterns of humanities data analysis, as well as the conflicts and challenges highlighted by these patterns, to bring some of their tacit assumptions into greater focus. The humanities offer an outstanding case study in the human factors of applied ML. Computation is simultaneously vital and foreign. On the one hand, scholarship has always required careful organization and detailed analysis of complex, difficult cultural materials. On the other hand, new technology lies far from the standard training of most humanities disciplines. Humanists therefore have strong motivation to use ML methods, but at the same time are unfamiliar with quantitative methods. They are, at the same time, experienced enough to develop creative new ways of using computational methods and also inexperienced enough that problems in tools and interfaces cannot be ignored. As a result, assumptions that would be unchallenged or unnoticed in other fields appear clearly. Specifically, humanists tend to assume a tool-building paradigm. For example, Graham et al. use the language of instrumentation in defining a "historian's macroscope" [14].

In humanities and social science research, researchers are rarely presented with a case in which a decision must be made, and when there are classifications they can be uncertain or debatable. In this chapter, we show that the same method—automated classification—can be used to illuminate complexities, identify interesting outliers, and call into question overly simplistic categorizations. "Mistakes" therefore become not a failure to find the truth but an opportunity to recognize the true complexity of a cultural artifact.

Identifying paradigm conflict. One of the areas where the tensions between the agent-building paradigm and the tool-building paradigm become clear is interpretability. Discussions about interpretability are fundamentally about values, and how those values are encoded into the functions that we seek to optimize. Is it enough to get the right answer, if you don't know *why* you got the right answer? Although interpretable models can reveal problems in data collection that result in overall improvements, in the short term predictive performance often trades off with interpretability [6].

Paradigm conflict can be hard to recognize because the technical details may be identical, but differ only in use. Mullainathan and Spiess [19] draw distinctions between methods that apply to the same model but focus on different

aspects of the model. For example, an agent-builder might use a linear regression to make predictions given features and never look at the parameters, while a tool-builder might use the exact same regression on the same training data, but pay attention to the sign and magnitude of the regression parameters to characterize effects, or use the residuals to identify interesting outliers. Charmichael and Marron [8] additionally point out that prediction and inference lie on a spectrum, and that sub-problems of one type may lie within larger problems of the other. Shmueli [28] draws a similar distinction between explanation and prediction, and argues for increased emphasis on prediction in scientific fields. He further highlights a dangerous tendency by which explanation is implicitly assumed to imply prediction.

In some cases users may apply both predictive and interpretive methods, but their relative importance is nevertheless defined by their value system. Methods for neural network interpretability such as LIME [24] are often presented as a way of providing post-hoc explanations or at least reassurances to users, who are still assumed to be mostly training independent agent-style models. The question of whether these methods are actually providing accurate and useful information remains open [16, 3]. In contrast, under the tool-building paradigm, interpretability is a vital, first-class proposition rather than a secondary, potentially valuable, but generally compromisable, attribute.

Interaction through Modeling Choices. So what is the interaction pattern of the tool-building paradigm? The process of applying an automated system to a data stream is filled with choices, from explicit parameter settings to implicit data pre-processing steps. Data streams in business and scientific processes can be complicated, but are often the product of existing computerized systems, and are—relatively—standardized. In contrast, much of humanities "data" is the result of centuries-old, idiosyncratic cultural processes that may be subject to centuries of additional preservation, adaptation, and curation before any digitization process. Humanities data is *weird* in ways that make it difficult to force into the specific format required by an algorithm. This is not to say that non-humanities data is not challenging and idiosyncratic, but cultural artifacts are, in our experience, more so.

Developing good tools for data curation and model interpretation is vital both to prevent incorrect use of computational methods and to encourage creative new uses. Humanities users often repurpose existing tools or use them in unexpected ways to suit their specific needs, and this creativity and nonstandard use is vital to making computational tools organically relevant in new fields. But poorly designed interfaces may make it possible for inexperienced users to use tools in ways that produce results that are irreproducible or overly sensitive to specific parameter choices. At worst, tools may silently allow researchers to perform experiments that cannot possibly produce meaningful outputs [11].

In this work we step through two case studies in the use of computational tools for humanistic inquiry, focusing on the stages of these processes that involve user interaction, often in ways that are hidden or poorly documented.

Our goal is to recast as *interactions* processes that are often described in terms that privilege a mathematical or computational view. While this interaction-centered approach is not new [28, 20, 5], many tutorials and methodological research articles focus almost exclusively on what the computation is doing. If there is discussion of what the user is doing, it is limited to a small side note about data preparation. Any findings are reported in a disembodied way, with no mention of who is doing the "finding" or how they did it.

Here we take the opposite approach, treating ML tools essentially as off-the-shelf packages and focusing on innovative ways to use them—much like the experience of humanities users. We focus on two case studies relating to collections of humanistic interest. These examples emerged in the course of our research on methodologies to support cultural analytics. We chose them because they demonstrate tool-building approaches, but also because in the process of presenting them to reviewers we experienced feedback that indicated a conflict between agent- and tool-building paradigms.

1.1 A Dadaist "reading" of Dada

In the first case study we explore the use of neural-network image classification tools to both identify and call into question the boundaries defined by art history. We use a classifier, but our goal is actually to question whether the images are classifiable.

The Dada movement was a community of avant-garde artists in the early 20th century. It emphasized a playful and irreverent aesthetic that deliberately blurred the boundaries of art, for example by presenting a porcelain urinal as a sculpture or by creating a poem by cutting a newspaper article into individual words and randomly shuffling them. Instead of physical newspapers we cut up digitized avant-garde periodicals from Princeton's Blue Mountain Project. Our initial corpus contains more than 2,500 issues from 36 different journals—over 60,000 pages in total. We use a convolutional neural network (CNN) to transform images into numerical vector representations.[1] These features are not readily interpretable to the human eye, but they may correspond to high-level concepts such as human faces, flowers, and fields of grass [33]. For each feature, we extract a number representing the feature's measured presence within an image—a large value indicates the feature is strongly detected, while a value near zero indicates its absence.

CNNs are powerful tools for analyzing images. Although the output of the final layer of a CNN will identify the object categories that it was trained to recognize, the output of the next-to-last layer has been shown to produce powerful, high-level visual features. These features are generic enough that they can be used by other image analysis systems [22]. By using these features as our computational cut-ups, we will in essence be asking what CNNs "see" when they look at Dada and more broadly the avant-garde.

We then use a simple ML classifier for two prediction tasks: whether each

[1] We use the ResNet50 model pretrained on ImageNet, which is available through Keras.

image contains music notation or not, and whether each image is *Dada* or *Not Dada*. Our purpose in this work is not to achieve high performance or to create a "good" classifier, but rather to use a relatively simple task (music identification) to provide users with intuitions about how CNNs "see" and to use a much more difficult task to get a new perspective on what visual features might distinguish a notoriously undefinable art movement from other avant-garde work. One of the comments we received from reviewers involved ideas to improve classification performance — a perfectly reasonable and expected request from an agent-building perspective. In this case, however, it was the *uncertainty* of the model itself that was of interest, as our goal is not to distinguish categories as well as possible, but to measure how difficult it is to distinguish categories.

1.2 Studying themes in science fiction

Our second case study is motivated by creating a tool to identify and quantify cross-cutting themes in a literary collection. Theme, genre, and subject in literature are all poorly-defined concepts, and are often the subject of overly simplified post-hoc narratives of literary history [30]. We wanted a system that could identify patterns at large scale that would be both recognizable but also potentially surprising.

Science fiction is both a well-established category of genre fiction and a category that is difficult to define by writers, readers, and scholars alike. Instead of asking what science fiction *is*, we ask what it is *about*. What topics are prominent within works of science fiction? And, how have these topics changed over time? We use topic modeling to identify a set of topics that are *grounded* in the texts of over 1,200 works of science fiction. Our working collection contains science fiction written by 245 authors[2] from the early 1800s to the present day. Since most of these works are currently protected by copyright, we use non-consumptive versions (i.e. page-level word counts) from the HathiTrust Research Center's Extracted Features Dataset [7]. In this case study, we use an LDA topic model [2] to generate thematic word clusters that we use as a tool to explore our large collection of science fiction.

Topic models generate a representation of a text collection in terms of "topics" (i.e. probability distributions over words) such that each document comprises a combination of topics. For tasks in the agent-building paradigm, these models are seldom used having long been superseded by large neural language models, which provide better performance for input-output-style objectives [4]. Nonetheless, topic models remain the most popular choice for tool-building approaches in the humanities because they produce highly *interpretable* thematic factorizations. Moreover, their representations estimate token-level topic assignments which naturally support grounded exploration of a collection. This lets users "see" a collection at multiple scales giving both a broad overview of the themes within a collection and an index of the specific documents and passages where these themes occur [5].

[2]We consider collaborations as separate authors.

Fundamentally, topic models are best for examining complexity rather than explaining it. Using a topic model is similar to conducting an archaeological survey, its purpose is to find and highlight where to dig further. In our science fiction example, we will use topic modeling to identify the topics—both expected and unexpected—that are present within large swathes of our collection to gain insights on what science fiction is about and how that has changed over time. These findings should not be considered a final product, but rather the groundwork necessary for refining research questions and analyses.

2 Data curation and preparation

For most humanists the primary context of interaction with ML methods is data preparation. In this section we describe the procedures we used for our two case studies. Dataset preparation is important and understudied. Most work in ML uses standardized, downloadable data sets because they are easy to use and ensure comparability of results. At the time of writing, for example, the GLUE benchmark of standard NLP tasks has over 3,300 citations since its publication in 2018 [32]. In contrast, when the goal is to use ML as a tool to study a collection, typically every study will involve the creation of a new collection and the adaptation of that collection to the format required by the chosen algorithm.

In discussing data preparation there is a subtle but important distinction between generating data in a *valid* format and generating data that produces *useful* results. Problems in the first case can be frustrating and time-consuming for researchers, who often have to diagnose cryptic error messages due to incorrect file formatting. The second case, however, can be even more concerning. Data may be in the correct format to allow a software package to run without errors, but may nevertheless contain artifacts, inconsistencies, or "shortcut" features [13] that produce low quality results.

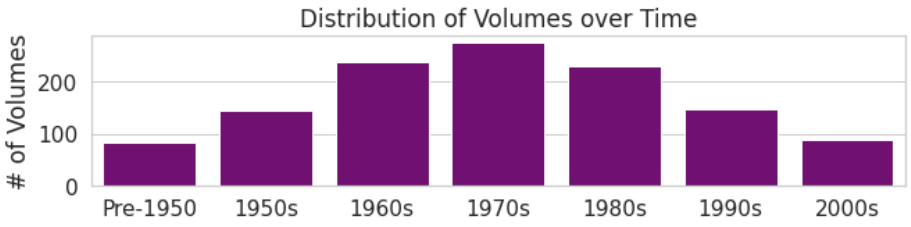


Figure 1: Our collection is not evenly distributed across time. The majority of texts were published after 1950 with the 1970s as the most prevalent decade.

2.1 Descriptive statistics

Interaction is critical in the early phases of a data-driven analysis in order to assess the consequences of the data collection processes that led to a data set being available. As data sets become larger and collection processes more automated, it becomes increasingly likely that a data set may be unbalanced in various ways that become clear through descriptive statistics. There is increasing support for this type of analysis, for example the Know Your Data project.[3]

As an example, our science fiction collection is not evenly distributed across time, as shown in Figure 1. Although our collection spans multiple centuries, including texts from the early 1800s to the 2000s, the vast majority of the collection was published in the 1950s or later. Based on the distribution of publication dates, we choose to represent all works published before 1950 as an equivalent "early" category rather than treating all decades equally. About half of these pre-1950 works are in the public domain (i.e. published before 1927). The remaining decades have enough representation to remain distinct. The 2000s has the fewest with 89 works, while the 1970s has the most with 275.

In our periodicals collection we had expected to find primarily visual art and literature, but there is also substantial content about music. The five periodicals *La Chronique musicale*, *Dalibor*, *Le Mercure* (*S.I.M.*), *Niederrheinische Musik-Zeitung*, and *La Revue musicale* are represented in the corpus by 1,405 issues and 27,791 pages. The majority of pages containing music notation come from these five journals. Using page-level metadata for each periodical issue, we identify 3,450 pages containing music notation.[4] Only ninety-one of these pages come from the thirty-one other periodicals. The availability of human labels for pages containing music notation and the frequency of these pages in the collection prompted our use of this feature as a proof-of-concept task, but we would not have been aware of either feature without inspecting the data set.

2.2 Segmentation

An important choice for users involves grounding abstractions and concepts in actual data. One of these abstractions involves defining the unit of analysis. For example, many text processing tools define a concept of a "document", but researchers in applied ML often struggle to map this abstract concept to specific examples, or fail to recognize the implications of these decisions [21]. To a researcher studying novels, a natural definition for "document" may be an 80,000 word novel, while to a researcher studying social media, a natural definition might be a 15 word microblog post. In both cases, definitions make sense in their context, but have radically different results when applied in a ML context. In particular, novels have been shown to be far too long and semantically variable for use as a "document" for LDA topic models [15]. Redefining the

[3] `https://knowyourdata.withgoogle.com/`

[4] We consider content marked as "Music" to represent musical content within a page. See `https://github.com/cwulfman/bluemountain-transcriptions`. Our research used transcripts accessed in May 2017.

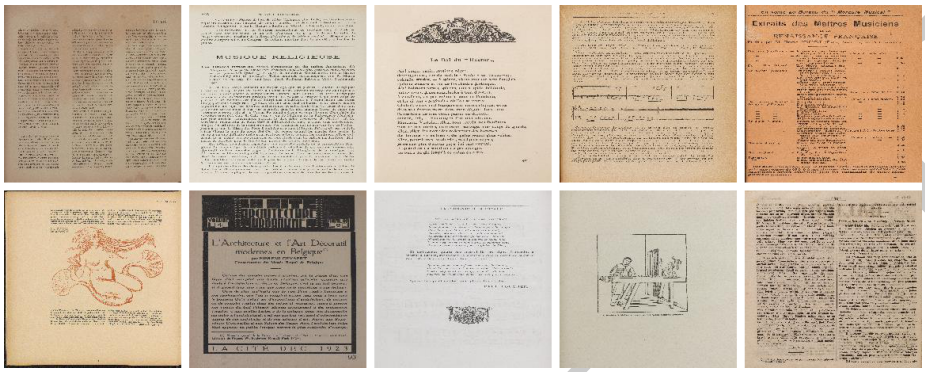


Figure 2: We resize images from full page scans to the thumbnail-sized square representations expected by the CNN. This figure shows ten randomly sampled pages.

abstract term to a more specific, quantitatively defined unit of text length such as paragraphs can have significantly better results [1]. For our science fiction collection, we operate at the level of pages. This is partially a function of the form our underlying data (i.e. page level word counts), but also that pages are generally a good unit for fiction since paragraphs of dialogue can be too short.

For images we choose to use the full page scan as it is present in the original scanned dataset. Alternatively, we could identify regions of page images and use those smaller, presumably more specific patches. But doing so comes at the cost of potential errors in image segmentation. In the case of avant-garde periodicals, we determined that the full pages were sufficiently visually coherent to avoid the additional burden of segmentation.

2.3 Modification

A second category of decisions that have significant effects is data modification. Images cannot be fed into off-the-shelf image models as-is. They must first be transformed into a format that these trained models expect. For our pretrained CNN model, this means we must shrink *and* deform our original images into small 224-by-224 pixel squares. This deformation will cause fine-grained details to be lost, but major elements such as layout, headers, and illustrations will generally be preserved. As seen in Figure 2, the images fed to the CNN remain recognizable but are similar to viewing the page from the far side of a room.

Other transformations of images might include cropping borders and background regions resulting from scanning print periodicals, rotation of images such as landscape-aspect artwork printed at a 90 degree angle to the text, or altering color. We found that certain periodicals, due to their physical shape relative to the scanner, tended to leave dark borders around the page image, which could provide a "shortcut" feature for an image classifier. Similarly, we evaluated both full color and grayscale images to determine whether artifacts resulting

from color printing processes could differentiate between journals.

In a similar way, we reduce the size of text documents by curating the vocabulary. Natural language is characterized by exponential relationships between the frequency of words: almost all words in a vocabulary are rare, but roughly half of all pages in our collection are made up of the 100 most frequent words. It is standard practice in topic modeling to remove high frequency words as they would otherwise "crowd out" more meaningful words, and to remove very low frequency words as it is difficult to gather meaningful information about them. There is substantial evidence that other similar modifications to the words that are present in documents can have predictable effects on modeling [12]. In general, our research indicates that many commonly used text preprocessing techniques such as stemming and aggressive stopword removal [26, 25] are less effective than many researchers believe, and should be avoided.

A more difficult problem arose in studying the science fiction collection: the names of characters and settings. Our goal is to find the prevalence of thematic elements in a large collection of novels and map their concentration over time. What we find, however, is that the patterns that are most accessible to the topic model inference algorithm are words that are characteristic of the specific imagined worlds of authors. This problem does not occur in more familiar news and scientific publication applications, where authors rarely invent unique vocabulary, and was surprising and unfamiliar to many text mining researchers. It was immediately familiar to those who had attempted to model novels, however, where the most prominent themes are almost always simply lists of character names and locations [15].

In one sense, producing clusters that strongly correspond to the work of a single author is the correct, optimal behavior for the algorithm: these words are frequent and only occur in specific contexts. But for our interactive purposes, these results do not provide a satisfying tool to identify and measure cross-cutting themes over decades. If we were interested in author signals, we wouldn't need ML, we would just look at metadata. Moreover, these authorial clusters may not be obvious and could lead to misinterpretations of the themes within a collection.

We can illustrate this problem by examining three topics produced by an LDA topic model (without any interventions). There is an Anne McCaffrey Dragonriders of Pern topic whose top terms (*f'lar lessa weyr robinton hold dragon f'nor lord dragons benden*) are clearly and exclusively names and settings which make it easy to flag as problematic. By examining the tokens assigned to this topic, we find that the vast majority come from Anne McCaffrey's works. She contributes over 122,000 tokens while the second largest authorial contribution is merely 400 tokens. A less obvious case, is an Isaac Asimov Robots topic whose top terms (*robot robots andrew human cully susan calvin brain being powell*) contain a mixture of character names and thematic looking terms. Because of the common terms *robot*, *robots*, and *human*, we might confuse this topic as a general one on artificial intelligence. When examining this topic's tokens, we find that Isaac Asimov's works are the primary contributors rather than the many works that focus on artificial intelligence. However, not all topics

with character names are bad ones. Our final topic (*sand pirx mars desert roger dust rock bass dunes crater*) seems problematic at first glance since it contains a mixture of common terms and character names like the Isaac Asimov Robots topic. But on closer inspect, both the common terms *and* the character names indicate that this is a topic about Mars. The names, although individually author specific, correspond to major characters from multiple stories by different authors, all set on Mars.

One possible solution to this problem could have been to modify the ML method to be aware of author signals and filter them out. Indeed, such modifications have been shown to be effective in isolating highly frequent words [31]. But from an interactive standpoint, doing so would have resulted in a special purpose algorithm that would add complexity for users while simultaneously introducing an opaque and inscrutable modeling layer. Instead, we chose to keep the solution within the framework of text curation. We use statistical tests to identify words that are significantly associated with particular metadata features such as authors [29]. We then calculate the proportion of instances of each word in each author's work that would need to be removed in order to be below a set threshold of statistical significance. Using these thresholds, we produce a new, modified copy of the text collection in which the statistically significant words have been randomly downsampled in each author's work to the point where they are no longer significantly associated. We are then able to use standard, well-tested LDA implementations to produce topics that are far less associated with individual authors, and better satisfy our interpretive goals.

2.4 Identifying duplicate instances

In addition to defining abstractions and applying data transformations, a third major category of decisions involves considering the *distribution* of features. Algorithms respond to patterns and quantitative relationships, but those quantifications may not reflect meaningful or desirable distinctions. One common example is the presence of exact or near-exact duplication.

Duplication is worrying from a computational modeling perspective because the goal of many ML methods is to identify patterns in data, or to identify regions of density in a data manifold [17]. Ideally, we want to recognize similarities between documents or images as a way of identifying more abstract relationships and clusterings. Duplication appears to fit this definition perfectly (indeed, too perfectly), and has the effect of "distracting" algorithms from potentially more meaningful and generalizable patterns. In the context of topic models, we find that duplicate documents effectively reduce the capacity of a model: if we train a model with K topics on a collection that has five copies of a single document, one topic is likely to be allocated entirely to representing that one document, so we are in effect unwittingly training a model with $K - 1$ topics [10, 27].

A more difficult and insidious form of duplication is near-duplicates, where there is some small variation that makes instances not exactly identical according to simple metrics, but close enough to have the same effect on models.

Computer vision techniques have been used to find modifications and reuses of images in Japanese woodblock prints [23]. Examples of image reuse in our image collection include advertisements and reproductions of the same artwork. In the science fiction collection there are texts with overlapping contents even after we remove multiple copies of the same work, due to compilations of previously published works, both shorter fiction and novels. For example, this collection contains *Shikasta* by Doris Lessing and the omnibus *Canopus in Argos: Archives* that contains it. In a few cases, near-duplicate works were included because of differing titles. The collection effectively contains two copies of Marge Piercy's *He, She and It* because this novel was published as *Body of Glass* outside of the US. Additionally since this collection includes collections (and accidentally a few anthologies), there may be content overlap across these entries. For example, it contains the Ray Bradbury collections *The Illustrated Man* and *The Stories of Ray Bradbury* which have 10 short stories in common. Since we are working with page-level word statistics rather than full texts we cannot perform classic text duplication techniques. However, we can check the similarity of word usage at the page (and volume) level.

3 Running models

While in some sense the most complicated part of the computational analysis workflow, running ML models, is often the simplest part from the user perspective. Choices available to users are often presented in the form of user settable parameters that are at best well documented and at worst at least enumerable. Unlike many data preparation steps, it is at least possible to list the options supported by a command line tool and verify what sort of values are applicable.

One category of decisions that needs to be made centers on parameters of the model itself that define the mathematical artifact that will be saved at the end of training. These include the dimensionality of the desired representation, such as the number of topics in a topic model and the number of dimensions in a vector representation. These determine the capacity of the model: a model with more dimensions will theoretically have more ability to encode patterns. Choosing values for these parameters can be confusing, and is often the subject of tutorials and online question-answering forums [5, 18]. In our experience, "how many topics should I use" is the single most commonly asked question by topic model users.

Another category of decisions involves parameters of the training *process*. These include learning rates, batch sizes, and the length of the training process. While the growth of general-purpose neural network libraries such as Keras [9] have increased the standardization of these parameters, they remain inscrutable to many users. The terms used in training can be confusing: "epoch" and "iteration" are used in unpredictable ways, many learning rate systems are deliberately given similar names (ADAM, AdaGrad, AdaDelta), parameters are named only as Greek letters. Other parameter names such as "dropout" and "warmup", while more descriptive than "beta" or "epsilon", remain inscrutable

even to experienced users. That said, these model parameters are at least discoverable by users, unlike many other data preparation steps described earlier that could have equal or greater effect but may require programming skills or additional software.

Finally, users need to be able to monitor the running of algorithms, which can take considerable time. Systems such as Tensorboard[5] can provide insight into problems in training. If data has not been prepared well or models are improperly specified, training may fail to improve the objective function or additional evaluation functions. For randomized inference that uses learning rates, a badly specified learning rate may cause objective values to oscillate between good and terrible values if they are too large or to remain unchanged if they are too small. Again, finding these problems and diagnosing them currently requires significant numerical optimization knowledge that is not expected for most users, especially in the humanities.

4 Model interpretation and analysis of results

The last major context of interaction between humanities users and models is the interpretation of results. In a more typical ML context this might include comparing numerical evaluation scores across model settings. In the humanities context, however, the goal of the analysis is to gain a new and different perspective on the data itself. Model interpretation therefore consists more often of "reading" the original collection through the model. For example, the *Neural Neighbors* project from the Yale Digital Humanities Lab[6] displays subsets of images from a collection of early photographs that have geometrically similar CNN vectors. In our avant-garde periodicals case study, we take a more model-focused view that centers on the distinction between categories as interpreted by the neural network representation.

4.1 Hits and misses: interaction through failure

While it is difficult to provide users with an interpretable perspective on the inner workings of a multi-layer ML model, we find that by focusing on correct and incorrect predictions we can give users a perspective on the kinds of features that these models are reacting to. The notion of true and false positives and negatives is familiar to most researchers and provides a framework to build a mental model of what the computer vision model does and does not respond to.

4.1.1 Proof of Concept: Seeing Music

Before testing whether a CNN can recognize Dada—a problem we do not necessarily believe is solvable—we demonstrate the feasibility of a simpler task:

[5] `https://www.tensorflow.org/tensorboard`
[6] `https://dhlab.yale.edu/neural-neighbors/`

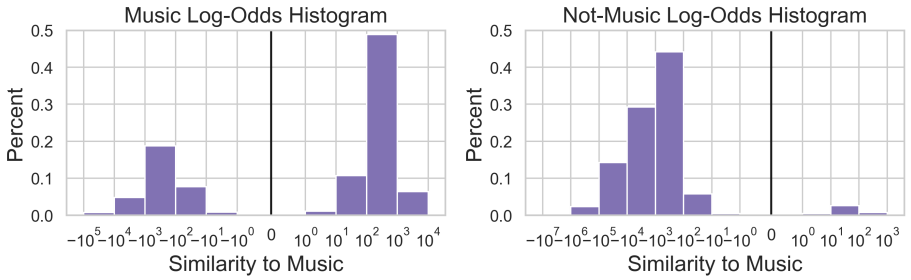


Figure 3: Histograms of prediction confidence for pages containing music (left) and pages without music (right). The classifier is more confident labeling pages as "Not-Music" no matter what the actual page type is.

identifying pages that contain music notation. Detecting music within our corpus is a relevant task, not only because music is an avant-garde art form, but because the Blue Mountain Project has a substantial number of music journals. It is fairly easy for a person to tell the difference between pages of musical scores and pages containing text and images, but how well will our CNN fare? If CNN representations do not distinguish between musical scores and paintings, it would be hard to trust their capability to distinguish Dada from Cubism.

We find that neural network embeddings are useful for recognizing pages containing music. The classifier makes mistakes that a human might not, but in ways that provide intuition about what it "sees." The classifier correctly labels 67% of the 3,450 pages with music as "Music" and 96% of the 55,007 pages without music "Not-Music." For each prediction, we can measure our classifier's confidence in terms of how much more likely it thinks a page should be labeled as "Music" rather than "Not-Music." Confidence scores with large magnitudes indicate a more confident classification, while a score's sign indicates its assigned label type. So, a large, positive confidence score indicates that the classifier is very confident that a page be labeled "Music." In Figure 3 we see that our classifier tends to be more confident when it labels a page as "Not-Music," even when it is wrong. This difference suggests that the vectors may better describe features associated with non-music page elements than music page elements.

To understand where the classifier goes wrong we compare the pages that are most confidently classified and misclassified for each label. It is important to emphasize that we are not interested in using the CNN in the agent-building style. We have perfectly good, high quality annotations for pages with and without music. The goal of the experiment and our interface is to provide users with insight into the workings of the model. In this case "mistakes" are not a sign of operational failure (we have no intention of using the model predictions) but rather a valuable signal that allows us to probe the ways in which CNNs "see" images differently from humans.

In Figures 4 and 5, we see that pages of sheet music are most confidently

recognized as "Music" and pages of tables are most confidently misclassified as "Music." These images share two prominent features: prominent horizontal lines and rectangular blank spaces. Given that the actual musical notes are poorly preserved in the deformed CNN inputs, it is reasonable that these are not the dominant visual features associated with pages containing music.

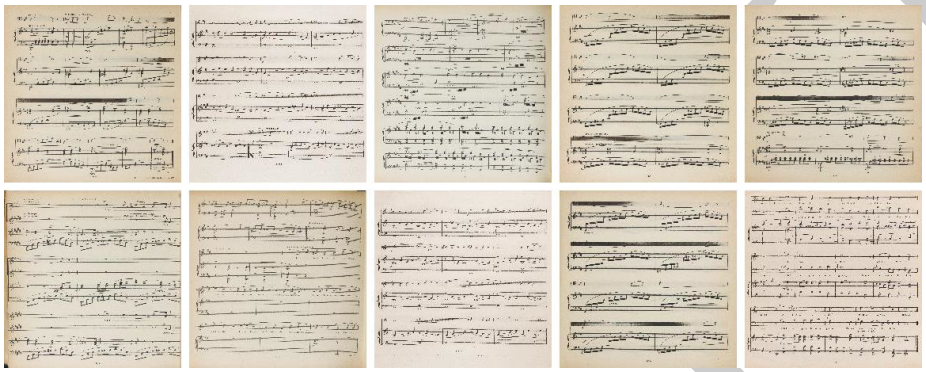


Figure 4: Ten pages most confidently, and correctly, classified as "Music."

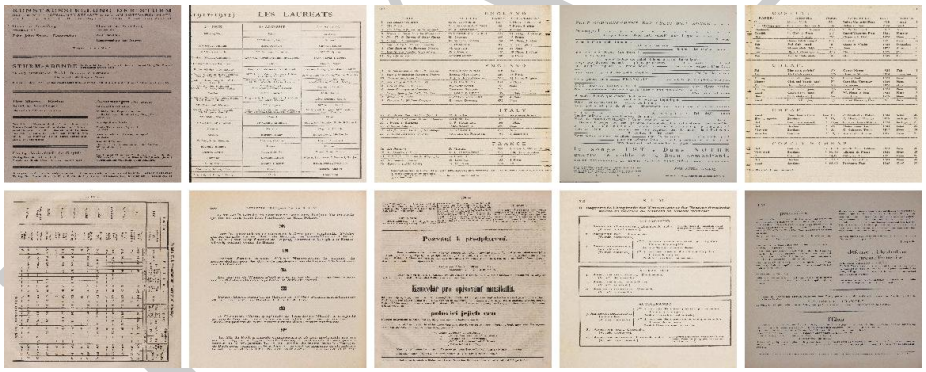


Figure 5: Ten pages most confidently misclassified as "Music."

Turning to the "Not-Music" label, we find color and pictures are the dominant visual features associated with pages without music. In Figure 7, we see that the top ten pages most confidently misclassified as "Not-Music" all contain pictures. Moreover, these pictures take up as much space within the page if not more than the musical elements. Many of these pages also include text.

Perhaps the most interesting of these confident "Not-Music" misclassifications is the bottom-right page in Figure 7, a scaled down image of a medieval folio. The rescaled CNN input image hardly looks like music, and, in a way, it is not. But looking at the original image in Figure 8, we see it does contain music, even though it looks nothing like modern musical notation. Additionally, the music is being seen through another medium: a picture, which could

Figure 6: Ten non-music pages most confidently classified as "Not-Music."

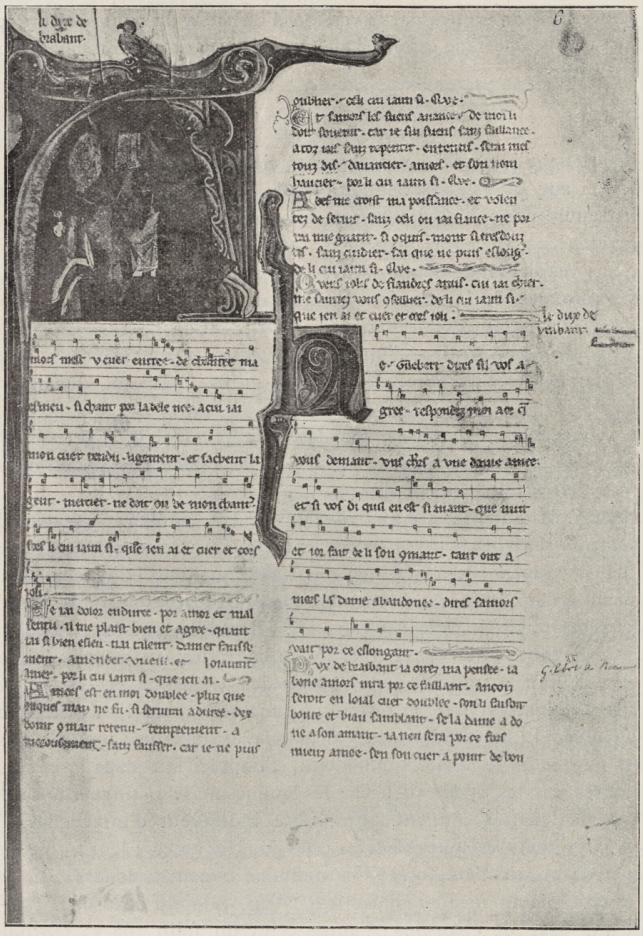
Figure 7: Ten pages most confidently misclassified as "Not-Music."

be misleading the classifier to the "Not-Music" label. From the perspective of a humanist, these "outliers" are potential sources of inspiration and further study, not problems to be removed.

From this interface, we can show that neural network embeddings are able to encode visual features that are useful for recognizing pages containing music. Users can infer from examples of correct and incorrect predictions that pages with music tend to have regular horizontal lines and rectangular white space, while pages without music tend to contain pictures and be in color.

4.1.2 Distinguishing Dada

Having established the effectiveness of our method, we can now turn to a more challenging task: whether we can distinguish "Dada" from "Not-Dada." Unlike the previous example, where there is a single, well-supported definition of pages with and without music notation, we are not confident that even Dadaists would be able or willing to distinguish Dada from other similar work. What then can we gain from this new computational perspective? We define labels at



PARIS, BIBLIOTHÈQUE NATIONALE,
ms. franç. 844, fol. 6, r°
(*en réduction*).

Figure 8: The contrast between model predictions and expert annotations can highlight surprising or unusual instances worthy of further study. This medieval manuscript is confidently misclassified as "Not-Music" but contains early musical notation.

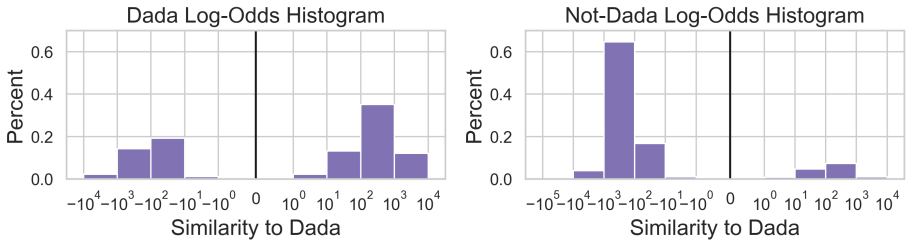


Figure 9: Histograms of prediction confidence for Dada (left) and not-Dada (right) pages. The classifier is more confident labeling pages as "Not Dada" no matter what the actual page type is.

the periodical level. For the purposes of this study, *Dada*, *291*, *Proverbe*, and *Le coeur à barbe* are "Dada" and all other periodicals are "Not-Dada." We acknowledge that this is a particularly coarse-grained perspective. A number of periodicals may feature works of Dada artists in specific issues, and these four periodicals might not always feature Dada artists, but these mistakes should have little effect on our classifier given the volume of actual not-Dada material.

We choose to exclude the five music journals from our analysis. Their sheer volume in the Blue Mountain Project would likely drown out the visual features that we are most interested in finding. Moreover, we want to avoid learning the naive feature that Dada does not contain sheet music. After this exclusion, we have 32,642 pages labeled "Not-Dada" and 182 labeled "Dada." Even with this removal, our labels remain extremely imbalanced. This makes classification more difficult, but nevertheless we can still examine what such a classifier "sees."

As expected, we find that neural network embeddings are less effective at distinguishing "Dada" from "Not-Dada," but they are significantly better than random. The classifier correctly labels 63% of the Dada pages and 86% of the not-Dada pages. In Figure 9, we see that the classifier is, as with music, more confident about its "Not-Dada" predictions. We speculate that other avant-garde movements may have visual signals that are easier to identify than Dada.

What then does the classifier "see"? When examining the classifier's most confident successes and mistakes in Figures 10–13, we find that the low-level features associated with Dada are high contrast, prominent edges, and the color red. In comparison, graded texture and photographs are considered not-Dada. From these low-level features, we see that abstract human forms are generally associated with Dada, while more realistic forms are not.

Figure 10: Ten Dada pages most confidently classified as "Dada."

Figure 11: Top 150 not-Dada pages most confidently misclassified as "Dada."

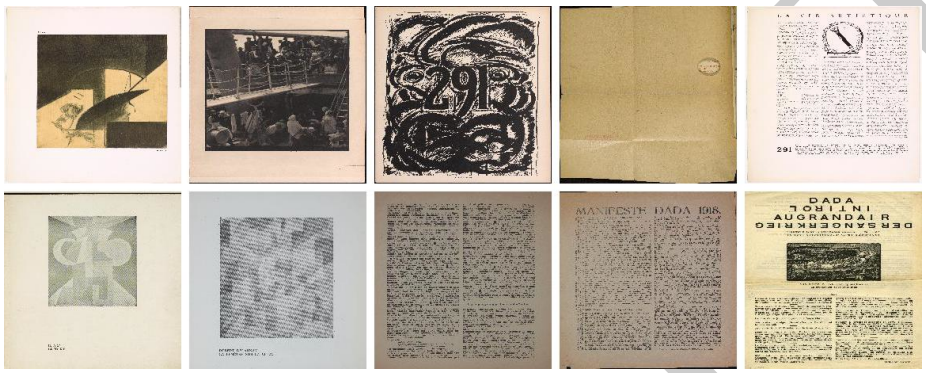


Figure 12: Ten Dada pages most confidently misclassified as "Not-Dada."

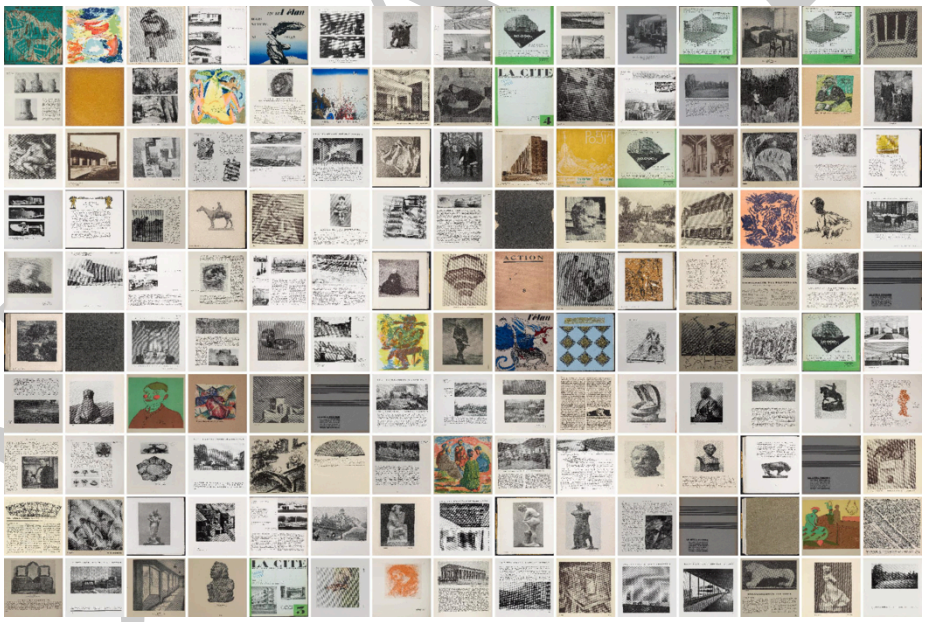


Figure 13: Top 150 not-Dada pages most confidently classified as "Not-Dada."

4.2 Exploration by Inspection

For our second case study we demonstrate a computationally assisted reading of a collection of science fiction texts using a topic model. Topic models are unusual as ML tools in that their primary, and arguably only, application is to provide human-interpretable interfaces to large text collections. Although topic representations have been applied to a variety of "downstream" tasks such as classification and clustering, such purposes are, in our experience, better solved by using either simpler word-counting representations or more sophisticated neural network representations. Topic models are a poor fit for the agent-building paradigm, but an excellent choice for a more interactive tool-building approach: they don't give you answers, they give you a different sort of text to be read.

Topic modeling provides a method for inspecting text at scale. It lets us measure things we already know—or at least suspect—about a collection more precisely, while also identifying things we *didn't* know which is arguably more important for exploration. In contrast to the agent-based paradigm that typically focuses on the most common patterns because of their likely importance for increasing performance, in the tool-based paradigm we often want to sift through the familiar to find the unfamiliar. Furthermore, topic modeling lets us map these broad patterns to specific instances in the text, which in turn gives us an opportunity for incorporating additional modes of reading.

In this case study we are interested in exploring the *content* of science fiction, and how these themes have or have not changed over time. We will examine both familiar and unfamiliar patterns identified by our topic model and examine how these patterns are distributed through our collection. The topic model detects and presents themes without human interaction *given the collection it is trained on*, but the specific content of the collection, and therefore the resulting topical patterns, are the result of the series of decisions that we made in interacting with the collection through pre-processing. The output of our topic model is far from a final answer to our posed questions, rather it gives us new insights on where to look further, a starting point to refine the questions we want to ask.

4.2.1 Where's the *science* in science fiction?

We begin our computationally-aided reading by looking for the familiar. What better place to start than to examine the topics that include the word *science*? We find three topics with *science* as one of their ten most frequent words: a general "science" topic (*work new* ***science*** *research scientists scientific knowledge project use problem*), an "academia" topic (*professor university college student students research school* ***science*** *work years*), and a "writing" topic (*story fiction* ***science*** *stories novel book writing has writer novels*). We suspect, but cannot confirm without access to the full texts in our collection, that the "writing" topic covers instances of "science fiction" rather than "science" on its own.

The mid-frequency "science" topic is found throughout our collection. More than two-thirds of volumes in our collection have over 50 tokens assigned to

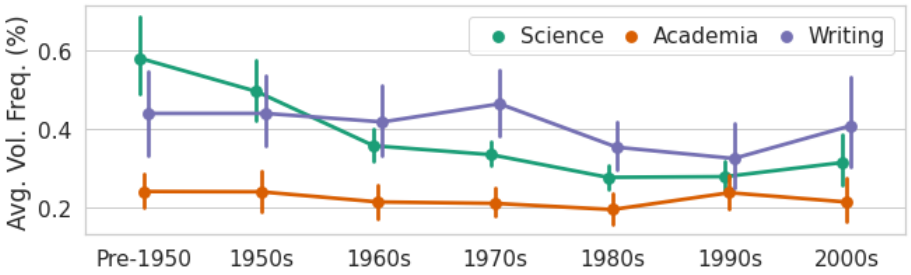


Figure 14: The general science discourse is used at high rates throughout the years but most prevalently in the 1950s and earlier. In contrast the "academia" and "writing" topics are used at similar rates across time.

this topic. Of these volumes, 143 have over 250 associated tokens, but no work has over 1,000 such tokens. Figure 14 shows that this topic is used throughout the years, but at much higher rates in the 1950s and earlier. The error bars—which correspond to 95% confidence intervals—indicate that these differences are significant, although the decrease could be a byproduct of our selection process. Still, this pattern suggests that although general science language is a common element of science fiction, it is most prominent in *early* science fiction. We suspect that later works use discourses on more specific emerging sciences and technologies that are captured by other topics within our model, but we would need to perform additional analysis to support this hypothesis.

In contrast, the lower-frequency "academia" topic is used at a similar rate throughout time. Only some 500 works in our collection have more than 50 tokens assigned to this topic and only 67 have over 250 such tokens. And yet, this topic is seen at higher volume-level concentrations with 20 works having more than 500 associated tokens (the general "science" topic only has 29 such works) and three with over 1,000 associated tokens. This difference in work-level concentration between the "academia" topic and the general "science" topic might be explained by their differences in specificity. The most prominent words of the "academia" topic can correspond to specific people (e.g. Professor James Dunworthy from *Doomsday Book* by Connie Willis) and places (e.g. Centauri University from *Empire Star* by Samuel R. Delany) within a narrative.

The "writing" topic, on the other hand, is associated with a particular type of publication: collections. Published compilations are more likely to contain added commentary on the writing of their collected works. Over half of the collection contains more than 50 tokens assigned to this topic, but 64 of these works contain over 500 associated tokens. Fifteen volumes have over 1,000 such tokens, of which all but two are collections. These two novels, *I Will Fear No Evil* by Robert A. Heinlein and *The Sword of Aldones* by Marion Zimmer Bradley, are editions with new (added) introductions.

4.2.2 The ships of science fiction

Since topic models can capture both the different meanings and contexts of a word, they can add depth and nuance to the questions we ask. For example, we might begin with a simple question of where (and when) spaceships are present in works of science fiction. We immediately find that there are no topics containing *spaceship* as one of their top twenty most frequent words, but there are six for *ship*. Of these six topics, four also have *space* as a top word and indeed relate to spacecraft. Two of these are mid-frequency topics: a "spaceship operation" topic (***ship space*** *control cabin pilot deck hull shuttle hatch aboard*) and a "space fleet" discourse (***ship*** *ships* ***space*** *aboard fleet crew planet our system captain*). These two contrasting views of spaceships suggest different narrative purposes: in the first, the ship as a physical, embodied location where a character can bang her knee on something, and in the second, a framework for an organization with an HR department. Reading with the model reminds us that "stories about spaceships" are not all the same, but rather vary in ways reflecting the goals of their authors. The other two low-frequency topics are more specialized: a "weapons" topic (***ship*** *beam toward energy fire power screen laser speed second*) and a "trade" topic (*van* ***ship*** *cargo port crew aboard trip* ***space*** *captain earth*).

In contrast, the non-space topics focus on "watercraft" (*boat water deck* ***ship*** *sea river boats wind sail shore*) and ship "command structures"(*captain* ***ship*** *crew officer sir men aboard cabin bridge chief*). As we might expect, the "watercraft" topic does not generally cooccur with the four space topics although there are exceptions (e.g. *The Earth Book of Storm Gate* by Poul Anderson and *Endymion* by Dan Simmons). Likewise, the "command structure" topic cooccurs with both spaceship and watercraft topics which speaks to both the topic's generality and the shared terminology used for spacecraft and watercraft. These additional topics highlight the connections between the ships of space and water. They are typically represented with similar command structures, but symbolize different environments and technology levels. This might leads us to push where these relationships break down, examining the cases that do not match our expectations.

4.2.3 What's in the background?

Viewing the collection through a computational tool draws our attention to "background" themes that might not otherwise be apparent. Stories contain many aspects not all of which we notice. While many people might predict that robotics and nuclear technology would be themes of science fiction, we find that a topic on news and media (*news story press television public newspaper people new paper read*) occurs at roughly the same frequency—and with much greater stability—than such expected topics. Over 500 works in our collection have 50 tokens assigned to this topic, but only 63 use over 250 tokens, and only one work, the novel *Steel Beach* by John Varley, contains over 1,000 tokens. This extreme usage can be explained by the novel's protagonist being a news-

paper reporter. Figure 15 shows little variation for this topic across the decades which suggests that news and media are a background element common to many science fiction works. We speculate that information sources may provide an important narratological structure that enables authors to convey information about unfamiliar worlds to readers. The model does not provide answers, but provides the beginning of questions, and an invitation to pursue the collection from a previously unexpected angle.

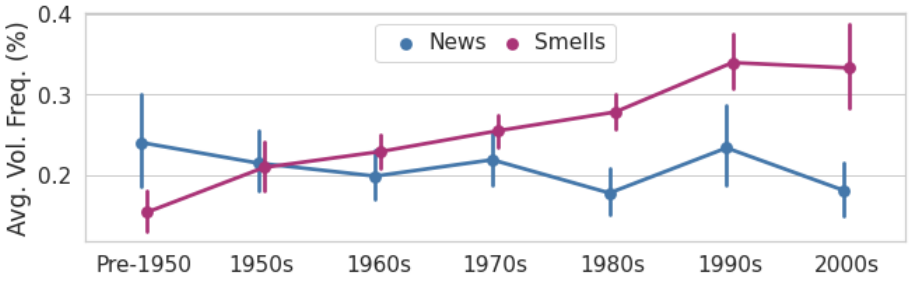


Figure 15: Discussion of news and media sources is a prominent theme throughout the collection, while smells are more prominent in the most recent decades.

4.2.4 Unexpected ~~finds~~ smells

Some topics surface themes that are less expected, but nevertheless meaningful. In our exploration, we were surprised to find a mid-frequency topic on smell (*smell air smelled odor scent nose nostrils smells breath taste*). Over half of the collection has over 50 tokens assigned to this topic, but only 80 use over 250 tokens. Interestingly, we find that this topic becomes more prevalent over time with a much higher average volume proportion for the 1990s and 2000s than in earlier years. This result could be suggestive for a researcher interested in how sensory impressions are represented in literature. Was there actually an increase in descriptions of scents in science fiction in the 1990s? And if so, what would it mean? Interacting with the collection through this model does not answer these questions, but it makes visible the subtle signals that might otherwise slip unnoticed to researchers reading one novel at a time.

5 Cyclical patterns in interaction

Practitioners commonly emphasize the cyclical nature of AI-assisted research. Iteration occurs for several reasons.

5.1 Parameter searches

When guidance on how to choose the value for user-settable parameters is undocumented or unknown, it may be necessary to enumerate a range of values and

try each one. With adequate computing resources, these searches can be done in parallel. Some parameters, such as learning rates and warmup steps, may be significant in making finicky optimization algorithms work for complicated models. Other parameters, such as the number of clusters or dimensionality of latent representations, may have more substantial effects on the representational capacity of models, not just their quality.

5.2 Multiple random seeds

A special case of parameter searches arises when we use randomized algorithms. In some cases algorithms are completely deterministic. Linear and logistic regression, for example, have a single global maximum that can always be reached from any starting setting. Other algorithms may be sensitive to a specific initialization, sequence of training examples, or randomized sampling within an optimization algorithm. These methods require a source of randomness, which is parameterized by an input integer, the random seed. In these cases, running the same algorithm with the same configuration but a different random seed can result in different behavior.

5.3 Mistakes

While rarely mentioned in publications, there are many opportunities for mistakes that destroy or invalidate results. A common error is incorrect file paths that lead to the inability to load data, or attempts to write to output directories that have not been created. Programs that have many output options may make it possible for users to forget to request that results are written to disk.

Problems with file paths are of particular concern. It has been frequently observed that current undergraduates have little experience with the abstraction of files that exist in a tree of directories. As cloud applications and storage become increasingly dominant in mainstream interfaces, and particularly in student-focused learning environments, it will become increasingly necessary to improve the design of functions involving reading and writing files.

5.4 Data problems

As the goal of humanities data analysis is to provide a new perspective on a data set, such approaches often reveal problems in data collection, digitization, and preparation. Unlike parameter searches and random seed variations, these problems are not clear a priori, so they may require repeating earlier modeling runs. As a result, they can cause frustration and unexpected delays.

5.4.1 Trouble with color

In the image analysis case study we noticed that color appeared to be a significant feature for both music and Dada. In the case of music, any saturated color indicated “Not-Music”, while for Dada the color red was more indicative

and with blues and greens for not-Dada. While sheet music is generally white, page coloring can vary due to paper and scanning quality. We want to verify that the non-color features perform well without the color cue, and see if the presence of pictures within a page remains the dominant "Not-Music" feature.

We found that removing color from CNN inputs had little effect on classification performance: 66% of pages with music and 97% of pages without music were correctly labeled. Additionally, the most confidently classified and misclassified images remained largely the same for each scenario except for correctly classified pages without music. This is what we had hoped to observe. It indicates that the classifier relies on features other than color. By removing color we also confirmed that the presence of pictures is an important feature for pages without music. As seen in Figure 16, pages containing pictures—both illustrations and photographs—are considered the least musical.

Given the prominence of red in "Dada" labeled pages, we were concerned that our results were overly dependent on this simple variable, and not able to generalize to shape or texture. We therefore reran the same analysis on grayscale images to measure the overall effect of color. The classifier's accuracy worsens for both label groups with resulting accuracies of 56% for "Dada" and 84% for "Not-Dada." Since this degradation is relatively small, we conclude that color is an important feature for distinguishing Dada, but not the only feature. We find that contrast, edge sharpness, and texture all remain prominent features for classification in grayscale.

It is perhaps unsurprising that color would play a role in distinguishing periodical groups, since page color is influenced by both content and printing method. If a journal has a distinctive page coloring, then it can easily be distinguished from other periodicals by this color alone. This feature can both cause pages with ambiguous content to be correctly identified and pages with otherwise highly similar content to be easily distinguished because of differences in color palette.

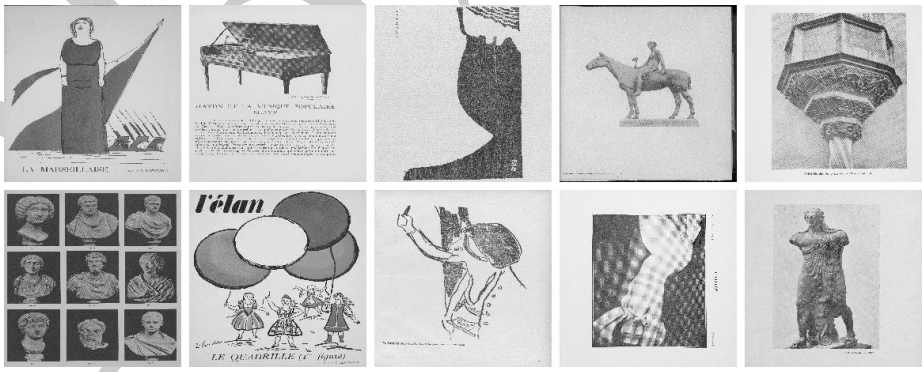


Figure 16: Ten grayscale pages correctly and most confidently classified as "Not-Music."

5.4.2 Persisting author correlation

Even though we are specifically modifying the collection to reduce the association of themes with specific authors, it is difficult to avoid the impact of prolific and focused ones. We find that there is still an Anne McCaffrey Dragon Riders of Pern topic (*lord hold between master queen star enough turns high good*) and an Isaac Asimov Robots topic (*human being law might must such without may robot beings*). Each author contributes the most tokens to their respective topic—over 50,000 tokens—with the largest volume-level contributions coming from books in their respective series. Despite the presence of these author-correlated topics, the topic model is also able to learn more general themes that previously were absent. Without modification there was neither a general, cross-cutting topic on robots nor one relating to dragons, but with the modification both exist (*machine robot machines robots human mechanical metal brain men built*; *bird cat wings cage birds dragon fly nest feathers egg*). So, while author-correlation still exists its harm is reduced.

These "bad", author-correlated topics might also be interesting to study in their own right. They highlight the common words that are used frequently and often uniquely by each book series. For example, in the Dragon Riders of Pern series, a *hold* is a fortified settlement and and dragons teleport by going *between*. Similarly, the Robot Series focuses on three *laws* of robotics that dictate the actions of *robots* and their interactions with *human beings*.

5.5 Expansion or refocusing of data sets

Sometimes an iterative step involves expanding or reducing a data set. Another result of gaining a better perspective on a data set is that it may become clear that the specific data set is either too broad or too narrow for the intended question, or that there are productive, alternative questions that would be better suited for a larger or smaller data set.

The most immediate steps for our science fiction example are to reduce the duplication and noise in the data set by removing anthologies and overlapping collections. Removing collections entirely would result in excessive curation since it would remove many novel-like works known as Fix-Ups that exist within this genre. Fix-ups contain previously published short stories that have been combined into a more novel-like form, thematically if not also structurally.

By construction the science fiction corpus was only meant to contain science fiction works. While this is less true in its actual form (e.g. science fiction writers write works in other genres), the collection could be expanded beyond the genre to better understand where boundaries blur. One reasonable expansion choice would be to include works broadly falling within the umbrella of speculative fiction. While speculative fiction is also difficult to pin down, it typically includes works from the genres of science fiction, fantasy, and horror. Instead of making these judgments directly, we recommend relying on external sources such as curated book lists.

We might also refocus our collection—and our subsequent analysis—by in-

corporating visual materials, namely the book covers of the included texts. A cover also conveys thematic elements, but not necessarily the same ones as the text itself. What is on a cover and what is *not* can provide a further dimension for organizing and comparing the works in our collection. What works that share many common themes will have dramatically different covers? What works will have very similar covers, but very different contents? This direction would shift the focus to the book as a multi-modal object—both textual and visual.

6 Conclusion

We presented two case studies that highlight the human interfaces in computational humanities projects. The *interaction* in this human-AI interaction is primarily realized in the process of collecting and curating datasets and in the process of analyzing model outputs. These case studies center the human actions in the project rather than the ML model. These projects leverage computational models for analyzing humanities collections—avante-garde periodicals and science fiction novels—but they are fundamentally focused on supporting the *human* interpretation of these collections.

In the Dada case study, CNNs are used to perform a deformative reading to provide a non-human perspective that allows us to question established categorizations and modes of thought. These perspectives are not always deep or meaningful, but even shallow and naive similarities encourage us to compare the seemingly incomparable. The process of trying to figure out "why did the model say that?"—even when it is wrong—forces us to see the familiar in new ways.

In the science fiction case, topic models are used to explore at a scale that is impossible to achieve through reading, but in a way that is clearly, transparently, and recognizably grounded in individual texts. Some aspects, like science and spaceships, are expected, but we see them in ways we might not have expected. Other themes like news media and sensory perceptions are less of a focal point, and are not what we might think to explore, but are equally present and contribute in meaningful ways.

These case studies follow the tool-building paradigm on machine learning, which differs fundamentally from the standard agent-building paradigm. They rely on the same tools and processes, but they have different needs and goals, and require different interaction patterns. Tool-building necessarily focuses on data curation and preparation of inputs, and analysis of outputs to a much greater extent than agent-building. The agent-building paradigm emphasizes the complexity and capacity of models and modeling, and thus inevitably incentivizes binaries and other simple, well-defined categorizations. Tool-building thrives on complexity and nuance to drive iterative exploration and refinement, as a way of generating meaningful insight and recognizing the true complexity of a phenomenon. Neither paradigm is wrong or right; they are simply different. They reflect different value systems and are both needed to move forward

the building and design of models and their applications. These developments, however, must begin by recognizing that these modes of interaction exist, are distinct, and equally worthy of study.

References

[1] M. Algee-Hewitt, R. Heuser, and F. Moretti. On paragraphs. scale, themes, and narrative form. *Stanford Lit Lab Pamphlets*, 1(10), 2015.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] T. Bolukbasi, A. Pearce, A. Yuan, A. Coenen, E. Reif, F. B. Viégas, and M. Wattenberg. An interpretability illusion for BERT, 2021.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] J. Boyd-Graber, D. Mimno, and D. Newman. Care and feeding of topic models: Problems, diagnostics, and improvements. In *Handbook of mixed membership models and their applications*. CRC Press Boca Raton, FL, USA, 2014.

[6] L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[7] B. Capitanu, T. Underwood, P. Organisciak, T. Cole, M. J. Sarol, and J. S. Downie. The HathiTrust Research Center extracted feature dataset (1.0), 2016.

[8] I. Carmichael and J. S. Marron. Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, 1(1):117–138, 2018.

[9] F. Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[10] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad. Redundancy-aware topic modeling for patient record notes. *PloS one*, 9(2):e87555, 2014.

[11] N. Z. Da. The computational case against computational literary studies. *Critical inquiry*, 45(3):601–639, 2019.

[12] M. J. Denny and A. Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189, 2018.

[13] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[14] S. Graham, I. Milligan, and S. Weingart. *Exploring Big Historical Data: The Historian's Macroscope*. World Scientific, 2015.

[15] M. L. Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[16] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[17] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

[18] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, 2018.

[19] S. Mullainathan and J. Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[20] D. Nguyen, M. Liakata, S. DeDeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, page 62, 2020.

[21] S. Passi and S. Jackson. Data vision: Learning to see through algorithmic abstraction. In *CSCW*, pages 2436–2447, 2017.

[22] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[23] J. Resig. Using computer vision to increase the research potential of photo archives. *Journal of Digital Humanities*, 3(2):3–2, 2014.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" explaining the predictions of any classifier. In *KDD*, pages 1135–1144, 2016.

[25] A. Schofield, M. Magnusson, and D. Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *EACL*, pages 432–436, 2017.

[26] A. Schofield and D. Mimno. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300, 2016.

[27] A. Schofield, L. Thompson, and D. Mimno. Quantifying the effects of text duplication on semantic models. In *EMNLP*, pages 2737–2747, 2017.

[28] G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[29] L. Thompson and D. Mimno. Authorless topic models: Biasing models away from known structure. In *COLING*, pages 3903–3914, 2018.

[30] T. Underwood. *Distant Horizons*. University of Chicago Press, 2019.

[31] H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. *Advances in neural information processing systems*, 22, 2009.

[32] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[33] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.