Chapter 4

Machine Learning in Digital Scholarship

Andrew Janco
Haverford College

Introduction

We are entering an exciting time when research on machine learning and innovation no longer requires background knowledge in programming, mathematics, or data science. Tools like RunwayML, the Teachable Machine, and Google AutoML allow researchers to train project-specific classification and object detection models. Other tools such as Prodigy or INCEpTION provide the means to train custom named entity recognition and named entity linking models. Yet without a clear way to communicate the value and potential of these solutions to humanities scholars, they are unlikely to incorporate them into their research practices.

Since 2014, dramatic innovations in machine learning have occurred, providing new capabilities in computer vision, natural language processing, and other areas of applied artificial intelligence. Scholars in the humanities, however, are often skeptical. They are eager to realize the potential of these new methods in their research and scholarship, but they do not yet have the means to do so. They need to make connections between machine capabilities, research in the sciences, and tangible outcomes for humanities scholarship, but very often, drawing these connections is more a matter of chance than deliberate action. Is it possible to make such connections deliberately and identify how machine learning methods can benefit a scholar's research?

This article outlines a method for connecting the technical possibilities of machine learning with the intellectual goals of academic researchers in the humanities. It argues for a reframing of the problem. Rather than appropriating innovations from computer science and artificial intelligence, this approach starts from humanities-based methods and practices. This shift allows us to work from the needs of humanities scholars in terms that are familiar and have recognized value to their peers. Machines can augment scholars' tasks with greater scale, precision, and reproducibility than are possible for a single scholar alone. However, only relatively basic and repetitive tasks can presently be delegated to machines.

This article argues that John Unsworth's concept of "scholarly primitives" is an effective tool for identifying basic tasks that can be completed by computers in ways that advance humanities research (2000). As Unsworth writes, primitives are "basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation." They are the building blocks of research and analysis. As the roots and foundations of our work, "primitives" provide an effective starting point for the augmentation of scholarly tasks.

Here it is important to note that the end goal is not the automation of scholarship, but rather the delegation of appropriate tasks to machines. As François Chollet recently noted,

> Our field isn't quite "artificial intelligence" — it's "cognitive automation": the encoding and operationalization of human-generated abstractions / behaviors / skills. The "intelligence" label is a category error. (2020)

This view shifts our focus from the potential intelligence of machines towards their ability to complete useful tasks for human ends. Specifically, they can augment scholars' work by performing repetitive tasks at scale with superhuman speed and precision. I proceed from this understanding to argue for an experimental and interpretive approach to machine learning that highlights the value of the interaction between the scholar and machine rather than what machines can produce.

***

Unsworth's notion "scholarly primitive" takes its meaning from programming and refers to the most basic operations and data types of a programming language. Primitives form the building blocks for all other components and operations of the language. This borrowing of terminology also suggests that primitives are not universal. A sequence of characters called a string is a primitive in Python, but not in Java or C. The architecture of a language's primitives changes over time and evolves with community needs. The Python and C communities, for example, have embraced Unicode as a standard to allow strings in every human language (including emojis). Other communities continue to use a range of character encodings, which grants greater flexibility to the individual programmer and avoids the notion that there should be a common standard.

For scholarship, the term offers a metaphor and point of departure. It poses a question: What are the most basic elements of scholarly research and analysis? Unsworth offers several initial examples of primitives to illustrate their value without a claim that they are comprehensive, including discovering, annotating, comparing, referring, sampling, illustrating, and representing. These terms offer a "list of functions (recursive functions) that could be the basis for a manageable but also useful tool-building enterprise in humanities computing." Primitives can thus guide us in the creation of computational tools for scholarship.

For example, with the primitive of comparison, a scholar might study different editions of a text, searching for similarities and differences that often lead to new insights or highlight ideas that would otherwise be taken for granted. As a tool, comparison can (but does not always) reveal new information. For an assignment in graduate school, I compared a historical calendar that showed the days of the week against entries in Stalin's appointment book. The simple juxtaposition revealed that none of Stalin's appointments were on a Sunday. This example raises questions for further investigation and interpretation. If Stalin was an atheist who worked at all times of the day and night, why wouldn't he schedule meetings on Sundays? Perhaps it was a legacy from Stalin's youth spent in seminary? Is there a similar pattern in other periods of Stalin's life? The craft of humanities research relies on many such simple initial queries. It should be noted that these little experiments are just the beginning of a research project. Nonetheless, the utility of comparison is clear. If anything, it seems so basic as to go unnoticed. This particular comparison offered an insight and new knowledge that led to further research questions.

Such beginnings are often a matter of luck. However, machine learning offers an opportunity to increase the dimensionality of comparisons. The similarities and differences between two editions of a text can easily be quantified using Levenshtein distance.[1] However, that will only capture the differences at the level of characters on a page. With machine learning, we can train embeddings that account for semantics, authors, time periods, genders and other features of a text and its contents simultaneously. We can quantify similarity in new ways that facilitate new forms of comparison. This approach builds on the original meaning and purpose of comparison as a form of "scholarly primitive," but opens additional directions for research and opportunities for insights. Rather than relying on happenstance or intuition to find productive comparisons, we can systematically search and compare research materials.

The second "scholarly primitive" that lends itself well to augmentation is annotation. This activity takes different forms across disciplines. A literary scholar might underline notable sections of a text by writing a note in the margins. A historian transcribes information from an archival source into a notebook. At their core, these actions add observations and associations to the original materials. Those steps in the research process are the first, most basic step, that connects information in a source to a larger set of research materials. We add context and meaning to materials that make them part of a larger collection.

When working with texts or images, machine learning models are presently capable of making simple annotations and associations. For example, named entity recognition models (NER) are able to recognize person names, place names, and other key words in text. Each label is an annotation that makes a claim about the content of the text. "Steamboat Springs" or "New York City" are linked to an entity called PLACE. Once again, we are speaking about the most basic first steps that scholars perform during research. I know that Steamboat Springs is a place. It's where I grew up. However, another scholar, one less versed in small mountain towns in Colorado, might not recognize the town name. They might identify it as a spring or a ski resort; perhaps a volcanic field in Nevada. The idea of "scholarly primitives" forces us to confront the importance of domain knowledge and the role that it plays in the interpretation of materials. To teach a machine to find entities, we must first explain everything in very specific terms. We can train the machine to use surrounding contextual information in order to predict — correctly — that "Steamboat Springs" refers to a town, a spring, or a ski resort.

As part of a project with Philip Gleissner, I trained a model that correctly identifies Soviet journal names in diary entries. For instance, the machine uses contextual clues to identify when the term Volga refers to the journal by that name and not to the river or the automobile. Where is the mention of "October" a journal name and not a month, a factory name, or the revolution? The trained model makes it possible to identify references to journals in a corpus of over 400,000 diary entries. This in turn makes it possible to research the diaries with a focus on reader reception. Normally, this would be a laborious and time-consuming task. Each time the machine predicts an entity in the text, it adds annotations. What was simply text is now marked as an en-

[1] Named after the Soviet mathematician Vladimir Levenshtein, Levenshtein distance uses the number of changes that would be needed to make two objects identical as a measure of their similarity.

tity. As part of this project, we had to define the relevant entities, create training data, and train the model to accomplish a specific task. This process has tangible value for scholarship because it forces us to break down complicated research processes into their most basic tasks and processes.

As noted before, annotation can be an act of association and linking. Natural language processing is capable of not only recognizing entities in a text, but also associating that text with a record in a knowledge base. This capability is called named entity linking. Using embeddings, a statistical language model can not only predict that "Steamboat Springs" is a town, but that it is a specific town with the record Q984721 in dbpedia. This association opens a wealth of contextual information about the place, including its population, latitude and longitude, and elevation. A scholar might have ample knowledge and experience reading literature — specifically, Milton. A machine does not, but it has access to context information that enriches analysis and permits associations. The result is a reading of a literary work that accounts for contextual knowledge. To be sure, named entity linking is not a replacement for domain knowledge. However, it is able to augment a scholar's contextual knowledge of materials and make that information available for study during research.

At this point, we are asking the machine not only to sort or filter data, but to reason actively about its contents. Machine learning offers the potential to automate humanities annotation tasks at scale. This is true of basic tasks, such as recognizing that a given text is a letter. It is also true of object recognition tasks, such as identifying a state seal in a letterhead or other visual attributes. A Haverford College student was doing research on documents in a digital archive that we are building with the *Grupo de Apoyo Mutuo* (GAM), of more than three thousand case investigations of disappeared persons during the Guatemalan Civil War. They noticed that many of the documents were signed with a thumbprint. The student and I trained an image classification model to identify those documents, thus providing the capability to search the entire collection of documents for this visual attribute. The thumbprints provided a proxy for literacy and allowed the student to study the collection in new ways. Similarly, documents containing the state seal of Guatemala are typically letters from the government in reply to GAM's requests for information about disappeared persons.

At present, several excellent tools exist to facilitate machine annotation of images and texts. Google's Teachable Machine offers an intuitive web application that humanities faculty and students can use to train classification models for images, sounds, and poses. To take the example above, the user would upload images of correspondence. They would then upload images of documents that are not letters.[2] Once training begins, a base model is loaded and trained on the new categories. Because the model already has existing training on image categories, it is able to learn the new category with only a few examples. This process is called transfer learning. For more advanced tasks, Google offers AutoML Vision and Natural Language, which are able to process large collections of text or images and to deploy trained models using Google cloud infrastructure. Similar products are available from Amazon, IBM, and other companies. Runway ML offers a locally installed program with more advanced capabilities than the Teachable Machine. Runway ML works with a wide range of machine learning models and is an excellent way for scholars to explore their capabilities without having to write code.[3] The accessibility of tools like

[2] In the Google Cloud Terms of Service there is specific assurance that your data will not be shared or used for any other purpose than the training of the model. More expert analysis may find concerns, and caution is always warranted. At present, there seems to be no more risk in using cloud services for ML tasks than there are for using cloud services more generally. See `https://cloud.google.com/terms/`.

[3] Teachable Machine, `https://teachablemachine.withgoogle.com/`; Google AutoML, `https://cloud.google.com/automl/`; RunwayML, `https://runwayml.com/`.

Runway allows for low-stakes experimentation and exploration. It is also a particularly good way for scholars to explore new methods and discover new materials.

For Unsworth, discovery is largely the process of identifying new resources. We can find new sources in a library catalog, on the shelf, or in a conversation. These activities require a human in the loop because it is the person's incomplete knowledge of a source that makes it a "discovery" when found. Given that machines reason about the content of text and images in ways that are quite unlike those of humans, machine learning opens new possibilities for discovery. When it comes to the differences in our own habits of mind and the computational processes of artificial networks, we may speak of "neurodiversity." Scholars can benefit from these differences, since the strengths of machine thinking complement our needs.

Machine learning models offer a variety of ways to identify similarity and difference with research materials. Yale's PixPlot, for example, uses a convolutional network to train image embeddings which are then plotted relative to one another in two-dimensional space with a stochastic nearest neighbor algorithm (t-SNE) (Duhaime n.d.).[4] PixPlot creates a striking visualization of hundreds or thousands of images, which are organized and clustered by their relative visual similarity. As a research tool, PixPlot and similar projects offer a quick means to identify statistically relevant similarities and clusters. This visualization reveals what patterns are most evident to the machine and provides a discovery tool for associations that might not be evident to a human researcher. Ben Schmidt has applied a comparable process to "machine read" and visualize fourteen million texts in the HathiTrust (n.d., 2018).[5] Using the relative co-occurrence of words in a book, Schmidt is able to train book embeddings. Schmidt's vectors provide an original way to organize and label texts based purely on the machine's "reading" of a book. These machine-generated labels and clusters can be compared against human-generated metadata. The value of this work is the human investigation of what machine models find significant in a collection of research materials. For example, with topic modeling, a scholar must interpret what a particular algorithm has identified as a statistically significant topic by interpreting a cryptic chain of words. The topic "menu, platter, coffee, ashtray" is likely related to a diner. In these efforts, Scattertext offers an effective tool to visualize what terms are most distinctive of a text category. In a given corpus of text, I can identify which words are most exemplary of poetry and which words are most exemplary of prose. Scattertext creates a striking and useful visualization, or it can be used in the terminal to process large collections of text.

Conclusion

As a conceptual tool, "scholarly primitives" has considerable promise to connect the intellectual goals of academic researchers in the humanities with the technical possibilities of machine learning. Rather than focusing on the capabilities of machine learning methods and the priorities of machine learning researchers, this method offers a means to build from the existing research practices of humanities scholars. It allows us to identify what kinds of tasks would benefit from being augmented. Using "primitives" shifts the focus away from large abstract goals, such as research findings and interpretive methods, to micro-methods and actions of humanities research. By augmenting these activities, we are able to benefit from the scale and precision afforded by

[4] See also `https://artsexperiments.withgoogle.com/tsnemap/`

[5] At time of writing, Schmidt's digital monograph *Creating Data* (n.d.) is a work in progress, with most sections empty until the official publication.

computational methods, as well as the valuable interplay between scholars and machines as humanities research practices are made explicit and reproducible.

References

Chollet, François. 2020. "Our Field Isn't Quite 'Artificial Intelligence' — It's 'Cognitive Automation': The Encoding and Operationalization of Human-Generated Abstractions / Behaviors / Skills. The 'Intelligence' Label Is a Category Error." Twitter, January 6, 2020, 10:45 p.m. `https://twitter.com/fchollet/status/1214392496375025664`.

Duhaime, Douglas. n.d. "PixPlot." Yale DHLab. Accessed July 12, 2020. `https://dhlab.yale.edu/projects/pixplot/`.

Schmidt, Benjamin. n.d. "A Guided Tour of the Digital Library." In *Creating Data: The Invention of Information in the American State, 1850-1950*. `http://creatingdata.us/datasets/hathi-features/`.

———. 2018. "Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries." *Journal of Cultural Analytics*, October. `https://doi.org/10.22148/16.025`.

Unsworth, John. 2000. "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" Paper presented at the Symposium on Humanities Computing: Formal Methods, Experimental Practice, King's College, London, May 2000. `http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html`.