The Page Image: Towards a Visual History of Digital Documents

Andrew Piper, Chad Wellmon, Mohamed Cheriet

Book History, Volume 23, 2020, pp. 365-397 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/bh.2020.0010

The Page Image

Towards a Visual History of Digital Documents

Andrew Piper, Chad Wellmon, and Mohamed Cheriet

On September 30, 1991, over 200 researchers assembled in Saint Malo, France, to convene the first ever conference on "document analysis and recognition."[1] The meeting brought together researchers from all over the world who for roughly the previous decade had been slowly changing the paradigm through which they approached the problem of the machinic understanding of the digitized page. Instead of thinking in terms of "characters" and "recognition," which underlay the long-standing field of Optical Character Recognition (OCR), they were gradually moving towards a more global and formal understanding of the page image as a whole. Researchers in the field of Document Image Analysis, or DIA as it came to be known, discarded the common assumption that the letter or the text was the ultimate referent of the bibliographic page. They focused instead on the heterogenous visual qualities of the page, or what they termed "the page image." "Document image analysis," writes George Nagy in a survey of twenty years of research in the field, is the "theory and practice of recovering the symbol structure of digital images scanned from paper or produced by computer."[2] DIA researchers turned the page image into an analytical object.

In moving away from a text-centric understanding of the page, research in Document Image Analysis offers an important new way of thinking about the bibliographic page that is different from what has traditionally been the case in computational approaches to studying culture, but that has deep roots in the fields of book history, bibliography, and textual studies. Whether in the guise of "natural language processing" (NLP), "optical character recognition" (OCR), or "text mining," computational approaches to pages have remained heavily influenced by a text-centric mentality, using the page image as an (often imperfect) means to an end, an object to be passed through rather than studied as something potentially meaningful in itself. At the same time, the fast-growing field of "image analytics," which ranges from facial detection to the analysis of newspaper illustrations, has largely maintained the text-image divide that has long dominated the study

of culture. Images are seen as independent of texts, whether as stand-alone objects or paratextual "illustrations." Emerging computational approaches to studying the past thus recapitulate long-standing disciplinary divisions and in the process reinforce textuality as the ideal object of study when it comes to documents.

Despite their numerous positive scholarly affordances, such text-centric approaches to the computational study of documents can constrain how we think about the past. Our "machine-readable" coverage of the past (as opposed to machine-observable), for example, is deeply biased in terms of both time and space. Currently, usable text data from digitized page images reliably stretches in any representative way only back into the nineteenth century, omitting well over two millennia of human writing. Similarly, while improvements are being made every day, OCR techniques still favor a very particular type of Roman-based font, which omits non-Western print traditions like Chinese woodblocks, non-print traditions like medieval manuscripts, or even regionally eclectic print traditions like German Fraktur.

Second, the text-centeredness of many computational methods and techniques obscures the layers of technological mediation that produce and make digital documents available in the first place. Ryan Cordell has argued that we need to think more about how each digitized edition or OCR'd version of an historical edition is another "setting" of that text, bound by a similar set of historical conditions under which the initial print (or manuscript) object was initially produced.[3] Like the particular printing press, house, and set of practices that governed the look and quality of a printed edition, OCR'd texts are similarly subject to particular machinery, institutional contexts, and human practices of correction and composition ("cleaning") that produce distinct outputs. Similarly, Matthew Kirschenbaum has been a vocal advocate for the physicality of born-digital documents, which are subject to the constraints of computing hardware.[4] And we have argued elsewhere that digitized page images should not be seen as universal and disembodied—available to everyone everywhere—but instead as physical items that can accrue histories of usage, circulation, and manipulation.[5] The page image is a thing that does things.

For a recent generation of bibliographers and book historians, such a reorientation around the visual and physical qualities of the page will come as no surprise. Bonnie Mak's *How the Page Matters* is a book-length study of the visual dimensions of the page, just as Christoph Windgätter's *Medienwechsel* is premised on a new visual "channelization" of writing during the Romantic era that had profound implications for practices of reading

and writing that came after.[6] Garrett Stewart's *The Look of Reading* foregrounded the visual qualities of writing in relief: in a pathbreaking pictorial history of writing's representation in the visual arts we see the ways in which artists and viewers acknowledge the challenging visuality of writing and reading.[7]

Such work draws on an important philosophical tradition that reframes the history of writing away from a phono-textual orientation towards a more visually inflected one. Rather than think of writing as a movement away from orality, as in the influential work of Walter Ong, philosophers such as Sybille Krämer have sought to draw attention to the visual evolution of writing and its epistemological consequences.[8] As Mara Mills has shown, when we do so we begin to attend to the integral relationship between reading, visual disability, and the evolution of media technology.[9] Similarly, thinking about the visual qualities of writing has been integral to the neurological study of human reading and cognitive disability and development.[10]

The computational framework of DIA thus shares many of the guiding assumptions that have shaped the field of book history as well other adjacent humanistic disciplines and even some sciences. And yet DIA and its methods and insights—despite important early contributions by Paul Fyfe and Natalie Houston—remain largely unknown outside the field and are rarely practiced within the computational study of culture.[11] In this essay, our aim is twofold. First, we want to introduce scholars of book history to the computational methods that can be applied to the study of books in a digital realm. The value of computational approaches to studying culture lies in their ability to significantly expand the scale of evidence considered when making inferences about the past. And yet a central challenge lies in the continued inscrutability of computational methods for scholars who are not familiar with them. In an effort to bridge such divides and to counter the tendency (on all sides) to set computational methods against more (purportedly) humanist ones, we describe the techniques and epistemic ideals of DIA, and highlight some of the continuities and divergences with the ideals, methods, and premises more typically associated with book history, bibliography, and philology. DIA, we would argue, is in important ways an unacknowledged outgrowth of such bibliographical and philological traditions. By describing DIA in some detail and pointing to some of its bibliographical and philological underpinnings, we hope to identify some of the affordances and limitations of the computational study of the past for a broader audience of book and cultural historians. At the same time, we also hope to convince scholars already using computational methods to study the past that

DIA has much to offer them. In particular, DIA can help them to reorient their thinking around the visual study of documents, a mode of thought that as we will show has a long and vibrant tradition within the field of book history. As others have argued, there is much to be gained by inserting the premises of book history into the digital study of the past.

To address these issues, we draw on our multi-year study of the history of scientific notation called "The Visibility of Knowledge." Bringing together the fields of book history, the history of science, and Document Image Analysis, we are interested in understanding the development of the graphic practices that accompanied and in many ways underwrote the production of scientific knowledge since the seventeenth century. As scholars like Adrian Johns, Elizabeth Eisenstein, Ann Blair, and a host of others have shown, the epistemic claims of scientific knowledge that emerged in the seventeenth century were intimately bound up with the medium of print.[12] The creation of new knowledge was a function not just of experiments and genial insights but of the organization and transmission of knowledge in printed form, one that relied on establishing new, and often visual, protocols of communication. Graphic practices like the use of footnotes, tables, diagrams, and figures were integral in establishing evidentiary norms even as the terminology and methods of scientific knowledge became more complex. Footnotes visually divided pages between text that referred to experiments or observations out in the world (the body), and text that pointed to more text, thereby creating new virtual communities.[13] Tables brought distinctly heterogeneous forms of information into dialogue with one another in two-dimensional spatial form (mirroring the page's own geometric orientation). Diagrams represented complex experimental or conceptual processes through new practices of visual synthesis.[14] And mimetic illustrations or "figures" oriented readers, and thus fellow scientists, to common objects and ways of seeing.[15] These visual elements were not merely supplementary to or popularizations of scientific knowledge. They helped constitute a unique visual language, one that has become a crucial component within the discovery, analysis, and defense of what continues to count as scientific knowledge.

Until recently, historians of science have typically relied on the manual sifting of documents. Even large, synoptic studies like that of Loraine Daston and Peter Galison's study of "objectivity" are uniquely constrained by the time and predilections of individual researchers.[16] To be sure, such fine-grained analysis of the documentary past has produced important insights. But computational methods not only allow us to observe much larger, and potentially more representative, swaths of the past. They also allow us

to test our assumptions against these more capacious collections. Instead of hand-selecting examples that confirm our prior beliefs, computational methods enable us to test claims of exemplarity about the past. They can help us better sample the past and allow us to give a better account of the error and potential biases of our own research.

"Scale" has, of course, emerged as one of the key concepts shaping research in the humanities in the past few years.[17] And yet our invocation of scale should not be seen as a slip into the rhetoric of transparency that either its proponents or critics often invoke—that more data is simply better. As historians and scholars across the humanities increasingly work with computational methods and colleagues from more computationally focused fields, it is crucial that we reflect on the computational conditions of knowledge as rigorously as we have the bibliographic conditions. For decades, book historians have taught scholars across the humanities how to better attend to the printed conditions of knowledge and the status of evidence in print archives. But we are only just beginning to theorize how the epistemic ideals and practices of scholarly knowledge are changing under the conditions of digitally remediated print archives. When we discuss the visibility of knowledge, then, we use the concept in two distinct but related ways: to refer to the object that we study and the means through which we study it, how knowledge is formed through seeing and how we come to see through knowledge.

In bringing document image analysis to bear on the history of scientific communication, one of our principal goals is to foreground the "page image" as a central unit of historical analysis. Independent of any particular findings that we may uncover over the course of our long-term project, our more immediate aim in this essay is to begin to take seriously the page image as an object of mediation in a double sense: to see the page *as* an image, that is, to focus on the page as a primarily visual rather than textual object and all of the qualities that attend its graphic identity; and second, to see the page image as an image *of* a page, that is, as a mediating object of knowledge rather than the thing itself. By combining the insights of book history and critical bibliography with the methodological insights of DIA, we hope to draw scholarly attention to the ways in which what we are seeing (before we begin to read or interpret documents) is first and foremost a representation of an absent artifact. In its most general sense, then, this essay orients us towards the study of the layered mediations of bibliographic knowledge in a digital environment. In doing so, we hope to offer another possible stance for relating to our printed past.

Seeing the Page from Bibliography to Machine Learning

Bibliographers and book historians have long reflected on the visual dimensions of pages. According to Philip Gaskell's *A New Introduction to Bibliography*, there are three primary dimensions to the visual qualities of pages (Table 1).[18] *Frames* refer to the use of borders, lines, or other elements that visually segment the page, including lexical units such as running headers or catchwords. *Letterpress* addresses a variety of issues related to the visual qualities of letters, from typeface to the quality of the rule to special letters such as swash or tailed letters. And *pieces* encompass aspects like headpieces, title pages, and colophons, where non-lexical visual elements have been used to decorate or ornament the page in segments.

Much of the initial emphasis on visual features within the field of bibliography was expressly guided by a principle of ornamentation. (Gaskell titled his chapter "Decoration and Illustration.") Visuality was seen as a decorative supplement to the more central issue of accounting for the production history of particular manifestations of individual works. It is telling that the study of the visual dimensions of pages takes up a tiny portion of Gaskell's handbook (roughly 6 pages in a 400-page book), indicating the marginality of the visual relative to other issues surround the study of the book.

Table 1.
Visual dimensions of the page according to Gaskell

Category	*Feature*
Frames	rules (lines)
	compartments (custom borders)
	frames (reusable borders)
	lexical (running headers, catchwords)
Letterpress	font
	typeface
	line endings
	swash letters
	tailed letters
Pieces	title pages
	headers
	running titles
	colophons

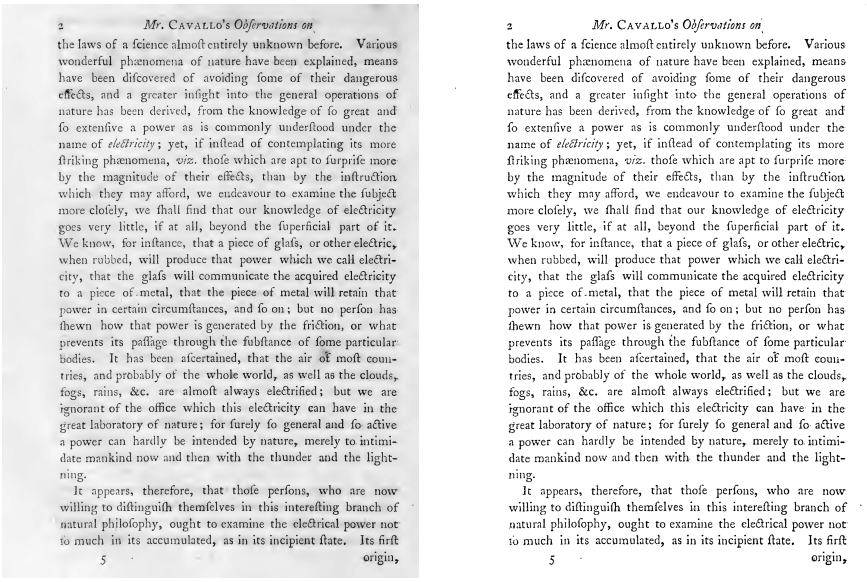
Alongside this bibliographic orientation to the decorative page, there is of course another longstanding strand of more art-historically informed research into the history of book illustration. Foundational works by John Harthan, Gordon Ray, Arthur Rümann, and Theodor Kutschmann have reconstructued the different periods and techniques of book illustration across different technological frameworks.[19] In this case, the visual is seen more as what Gerard Genette would call a "paratext," as something that resides next to, but is distinct from, the text proper. The visual is not associated with the page, but is treated as a distinct imagistic practice with its own set of conventions, genealogies, and practitioners.

More recent research has emerged in the previous decade that is informed by a theory of intermediality. This work has emphasized instead the ways that texts and visual features of books interact with one another, in turn shaping readers' interactions with books.[20] The aim of this work is to move past dichotomies between text and image and see the ways that texts function imagistically and images can be "read" textually. The visual is not seen as incidental to the textual in this tradition —whether as a form of adornment or as a privileged outside of the text—but as an integral aspect of the history of books.

We would argue that Document Image Analysis draws implicitly from all three of these traditions in different ways. Understanding this complementarity—however implicit or unacknowledged—is key to understanding the limitations and affordances that computational approaches offer to researchers in the humanities who study digital documents.[21] On the one hand, DIA methods are premised on the idea of *normalization* common to much data-driven research. Before analyzing the features of a page image, DIA researchers undertake a series of steps designed to separate and then isolate any effects of the imaging process from the visual qualities of the underlying document.

For example, one of the most essential steps in the process is that of *binarization*, which is used to differentiate what DIA researchers call the foreground and background of a page image (Fig. 1). Qualities that are relevant to the process of binarization can include the discontinuous color of the page through weathering (Fig. 1, left) or the bleeding of ink either through the page or across to an adjacent page (Fig. 2). Marks introduced through the process of imagization are also relevant to this step, such as the presence of dust or scratches, as when the object digitized is a microfiche and not the actual book. Another key step is known as *deskewing*, which corrects the distortions introduced by a scanner imperfectly aligned with the underlying

page (Fig. 3). Skew correction typically involves the creation of a histogram of the horizontal projection of the page's lines. Think of running a horizontal ruler across the page, one pixel at a time. The more black pixels of type you cross, the higher the bar on the right will be. Very short bars (or no bars) indicate the absence of black pixels (i.e. white space). Very high bars indicate a lot of black pixels, i.e. dense spaces of letters. The more skewed the page image, the flatter the histograms will be because as the type spreads or curves it looks wider (Fig. 3, left). A pointier, more highly differentiated histogram thus indicates a more evenly ruled page (Fig. 3, right).



2 *Mr.* CAVALLO's *Obſervations on*

the laws of a ſcience almoſt entirely unknown before. Various wonderful phænomena of nature have been explained, means have been diſcovered of avoiding ſome of their dangerous effects, and a greater inſight into the general operations of nature has been derived, from the knowledge of ſo great and ſo extenſive a power as is commonly underſtood under the name of *electricity*; yet, if inſtead of contemplating its more ſtriking phænomena, *viz.* thoſe which are apt to ſurpriſe more by the magnitude of their effects, than by the inſtruction which they may afford, we endeavour to examine the ſubject more cloſely, we ſhall find that our knowledge of electricity goes very little, if at all, beyond the ſuperficial part of it. We know, for inſtance, that a piece of glaſs, or other electric, when rubbed, will produce that power which we call electricity, that the glaſs will communicate the acquired electricity to a piece of metal, that the piece of metal will retain that power in certain circumſtances, and ſo on; but no perſon has ſhewn how that power is generated by the friction, or what prevents its paſſage through the ſubſtance of ſome particular bodies. It has been aſcertained, that the air of moſt countries, and probably of the whole world, as well as the clouds, fogs, rains, &c. are almoſt always electrified; but we are ignorant of the office which this electricity can have in the great laboratory of nature; for ſurely ſo general and ſo active a power can hardly be intended by nature, merely to intimidate mankind now and then with the thunder and the lightning.

It appears, therefore, that thoſe perſons, who are now willing to diſtinguiſh themſelves in this intereſting branch of natural philoſophy, ought to examine the electrical power not ſo much in its accumulated, as in its incipient ſtate. Its firſt

5 origin,

2 *Mr.* CAVALLO's *Obſervations on*

the laws of a ſcience almoſt entirely unknown before. Various wonderful phænomena of nature have been explained, means have been diſcovered of avoiding ſome of their dangerous effects, and a greater inſight into the general operations of nature has been derived, from the knowledge of ſo great and ſo extenſive a power as is commonly underſtood under the name of *electricity*; yet, if inſtead of contemplating its more ſtriking phænomena, *viz.* thoſe which are apt to ſurpriſe more by the magnitude of their effects, than by the inſtruction which they may afford, we endeavour to examine the ſubject more cloſely, we ſhall find that our knowledge of electricity goes very little, if at all, beyond the ſuperficial part of it. We know, for inſtance, that a piece of glaſs, or other electric, when rubbed, will produce that power which we call electricity, that the glaſs will communicate the acquired electricity to a piece of metal, that the piece of metal will retain that power in certain circumſtances, and ſo on; but no perſon has ſhewn how that power is generated by the friction, or what prevents its paſſage through the ſubſtance of ſome particular bodies. It has been aſcertained, that the air of moſt countries, and probably of the whole world, as well as the clouds, fogs, rains, &c. are almoſt always electrified; but we are ignorant of the office which this electricity can have in the great laboratory of nature; for ſurely ſo general and ſo active a power can hardly be intended by nature, merely to intimidate mankind now and then with the thunder and the lightning.

It appears, therefore, that thoſe perſons, who are now willing to diſtinguiſh themſelves in this intereſting branch of natural philoſophy, ought to examine the electrical power not ſo much in its accumulated, as in its incipient ſtate. Its firſt

5 origin,

Figure 1. Left, Original scanned image of page 2 from the *Philosophical Transactions of the Royal Society* (1778). Right, binarized version of the same page. Source: The Biodiversity Heritage Library.

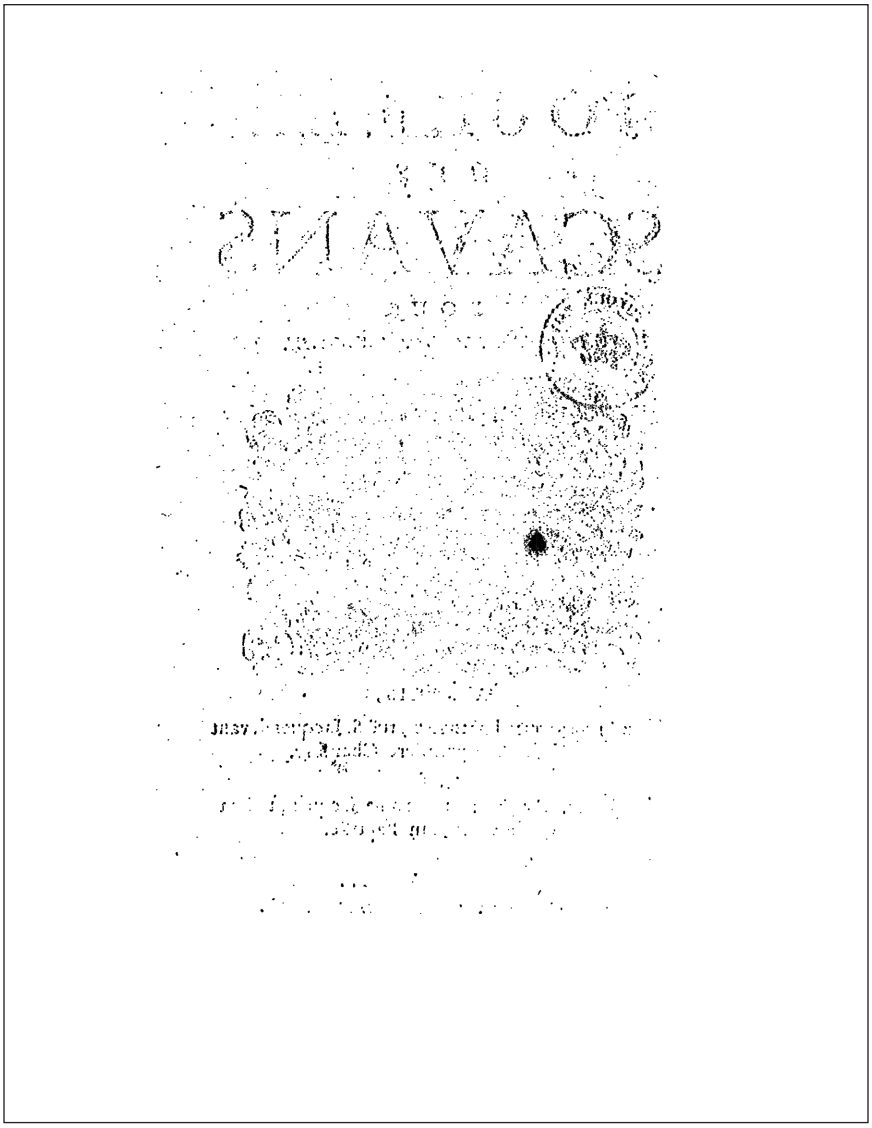
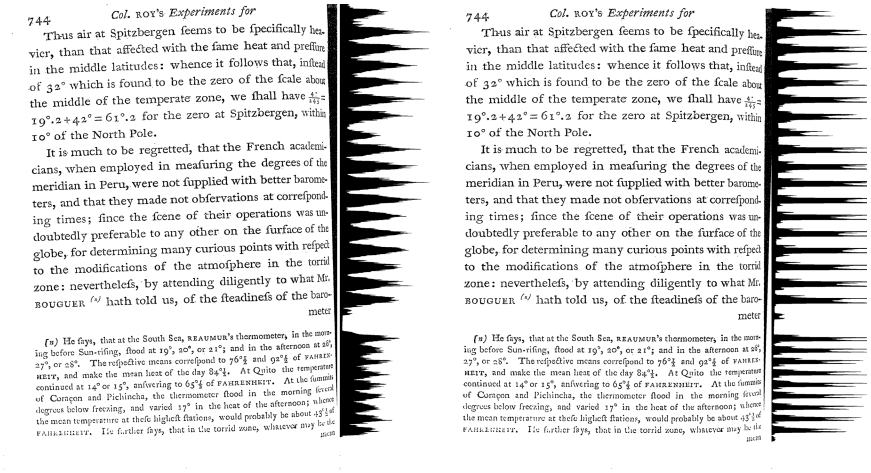


Figure 2. Example of unintentional typographic marks left through bleeding. Opposite title page, *Le Journal des Sçavans* (1684). Source: Bibliothèque nationale de France.


744 *Col.* ROY's *Experiments for*

Thus air at Spitzbergen ſeems to be ſpecifically heavier, than that affected with the ſame heat and preſſure in the middle latitudes: whence it follows that, inſtead of 32° which is found to be the zero of the ſcale about the middle of the temperate zone, we ſhall have $\frac{42}{19}$ = 19°.2+42° = 61°.2 for the zero at Spitzbergen, within 10° of the North Pole.

It is much to be regretted, that the French academicians, when employed in meaſuring the degrees of the meridian in Peru, were not ſupplied with better barometers, and that they made not obſervations at correſponding times; ſince the ſcene of their operations was undoubtedly preferable to any other on the ſurface of the globe, for determining many curious points with reſpect to the modifications of the atmoſphere in the torrid zone: nevertheleſs, by attending diligently to what Mr. BOUGUER (*u*) hath told us, of the ſteadineſs of the barometer

(*u*) He ſays, that at the South Sea, REAUMUR's thermometer, in the morning before Sun-riſing, ſtood at 19°, 20°, or 21°; and in the afternoon at 26°, 27°, or 28°. The reſpective means correſpond to 76°½ and 92°½ of FAHRENHEIT, and make the mean heat of the day 84°½. At Quito the temperature continued at 14° or 15°, anſwering to 65°½ of FAHRENHEIT. At the ſummits of Cotopaxi and Pichincha, the thermometer ſtood in the morning ſeveral degrees below freezing, and varied 17° in the heat of the afternoon; whence the mean temperature at theſe higheſt ſtations, would probably be about 43°½ of FAHRENHEIT. He further ſays, that in the torrid zone, whatever may be the mean

744 *Col.* ROY's *Experiments for*

Thus air at Spitzbergen ſeems to be ſpecifically heavier, than that affected with the ſame heat and preſſure in the middle latitudes: whence it follows that, inſtead of 32° which is found to be the zero of the ſcale about the middle of the temperate zone, we ſhall have $\frac{42}{19}$ = 19°.2+42° = 61°.2 for the zero at Spitzbergen, within 10° of the North Pole.

It is much to be regretted, that the French academicians, when employed in meaſuring the degrees of the meridian in Peru, were not ſupplied with better barometers, and that they made not obſervations at correſponding times; ſince the ſcene of their operations was undoubtedly preferable to any other on the ſurface of the globe, for determining many curious points with reſpect to the modifications of the atmoſphere in the torrid zone: nevertheleſs, by attending diligently to what Mr. BOUGUER (*u*) hath told us, of the ſteadineſs of the barometer

(*u*) He ſays, that at the South Sea, REAUMUR's thermometer, in the morning before Sun-riſing, ſtood at 19°, 20°, or 21°; and in the afternoon at 26°, 27°, or 28°. The reſpective means correſpond to 76°½ and 92°½ of FAHRENHEIT, and make the mean heat of the day 84°½. At Quito the temperature continued at 14° or 15°, anſwering to 65°½ of FAHRENHEIT. At the ſummits of Cotopaxi and Pichincha, the thermometer ſtood in the morning ſeveral degrees below freezing, and varied 17° in the heat of the afternoon; whence the mean temperature at theſe higheſt ſtations, would probably be about 43°½ of FAHRENHEIT. He further ſays, that in the torrid zone, whatever may be the mean

Figure 3. Example of a skewed page, left, with its corrected version, right. Notice how the histogram of the projected lines of the page changes from more uniform on the left to more differentiated on the right as it is deskewed. Source: Eighteenth-Century Collections Online.

Steps like binarization or deskewing highlight the distinction between what researchers believe to be the salient or intentional information on the page and the accidental or background information introduced through the process of imagization (what computer scientists might call "noise"). Book historians will immediately hear in this language echoes of the distinction bibliographers have long made between "intentional" and "accidental" marks, where accidental marks are errors or peculiarities to the page that obfuscate the "true" or intended information of the page. In this sense, DIA seeks an "ideal" version of an underlying work, one that is independent of the accidents of history. Just as earlier forms of bibliographic study sought to reconstruct an ideal version of a work in order to recreate an author's original intentions, DIA's attempt to normalize documents also idealizes them and, thus, makes numerous otherwise heterogenous documents commensurable with one another. Like the process of collating different editions to arrive at an authoritative, ideal text, the steps of normalization remove the accidents of historical production—whether it be the original printing process or the subsequent process of digitization—in order to recover the "original" document and its "true" features. It removes portions of the historical record, the reproduction history of texts, to arrive at a broader understanding of the historical record, in this case the production history of texts.

This distinction between accidental and intentional marks has, of course, been subject to numerous critiques in book history and bibliography. Since the work of D.F. McKenzie, Jerome McGann, and others, the sociological turn within bibliography has focused less on reconstructing intentions and ideal versions and more on accounting for the particularity of multiple editions and the different historical contexts in which they assume meaning.[22] In its attempt to account for more capacious numbers of documents—arguably the fundamental affordance of computational methods—DIA also follows within this sociological tradition by not reducing different versions of texts into single representations. It allows us to account for the differences and distinctions between large numbers of documents and, thus, gain a better understanding of the social context in which they were circulating. Second, there is nothing intrinsic to the qualities associated with "foreground" and "background" in DIA. In other words, what one DIA researcher treats as background—skewed images, distortions of dust from scanning microfiche, uneven rule in the printing process—another can treat as foreground. The basic categories of the process of normalization are defined by researchers, rather than by some pre-existing criteria. The prevalence of skew or the quality of a reproduction could be used to tell us something about the process of digitization itself, much in the way Ryan Cordell has argued that we ought to see OCR as a form of machinic typesetting.[23] Alongside the local idealizations that DIA makes in order to normalize documents to make them comparable at large scale, there is also an epistemological flexibility that is one of its greatest affordances. DIA can be seen in this sense as an ideal tool for book historians to describe and understand the different layers of technological mediation which constitute any particular digital document.

We label the second set of procedural steps used in DIA *analysis*. This step has a much clearer connection with more recent hermeneutic work in the field of book history (see Table 2 for a summary). Once the page has been "normalized," researchers then move to the analysis of particular features of interest. The first step in the analytical process is very often one of *segmentation*. Segmentation assumes that there are multiple qualities that belong to a page image and that these qualities have geometric or regional properties. When researchers focus exclusively on optical character recognition (OCR), the primary segments of the page are the *character* and the *line*. These can be identified using the method of "connected components" (Fig. 4) and "horizontal projection" discussed above (Fig. 5). A connected component is a continuous shape with no interruption to the flow of black pixels. In theory this captures individual letters, but in practice, especially

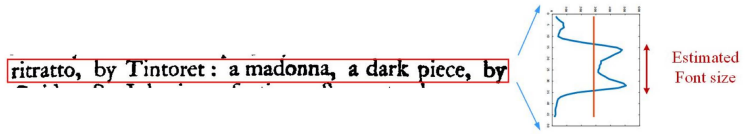
with historical documents and imperfect image scans, this approach can be subject to error. A line is then estimated as the range of the densest black pixels, with segments drawn between lines based on the range of the projected histogram (Fig. 5). But a segment may also be considered to be an illustrative dimension of a page, as when a page contains a decorative headpiece or an illustration proper. In this regard, DIA allows researchers to identify and focus on decorative and illustrative traditions that are well-aligned with the art historical vein of book history that we mentioned above.

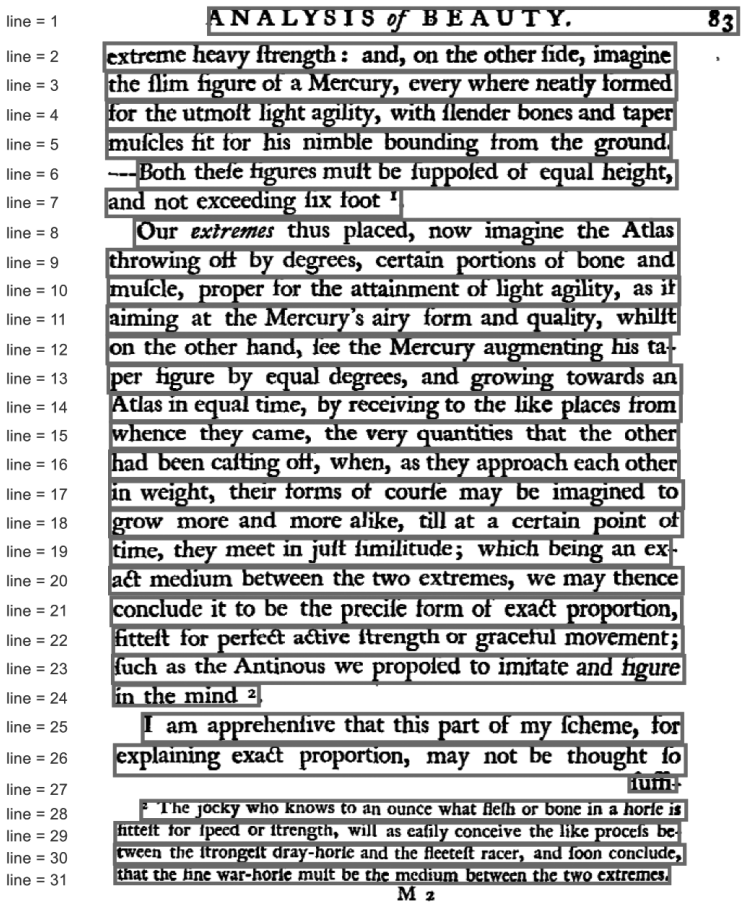
Figure 4. The grey boxes here represent connected components. In this example of "slender" from Hogarth's *Analysis of Beauty*, we see how the first connected component spans more than one letter due to the typeface. Source: Eighteenth-Century Collections Online.

As DIA has developed, however, researchers have increasingly tried to account for the visual heterogeneity of the page by focusing on segmenting different visual fields or what they call "views." The character and the line become parts of a larger set of possible page segments. According to this theory, a page image cannot be fully described by a single bibliographic formula. Each page requires a multiplicity of potential points of view. As Andreas Dengel and Faisal Shafait argue, in order to capture the "structure of complex documents," researchers have developed generalized models that

> represent a document as a set of layout or geometric structures. The set of layout structures is a collection of views such that each view represents a different layout interpretation of the document. Each layout structure itself is a set of geometric document objects and a set of geometric relations among them.[24]

The relationships between these multiple views and their identification by type (headers, footers, illustration, decorated letters, white space, borders,





ANALYSIS *of* BEAUTY. 83

extreme heavy ſtrength: and, on the other ſide, imagine the ſlim figure of a Mercury, every where neatly formed for the utmoſt light agility, with ſlender bones and taper muſcles fit for his nimble bounding from the ground. ---Both theſe figures muſt be ſuppoſed of equal height, and not exceeding ſix foot [1].

Our *extremes* thus placed, now imagine the Atlas throwing off by degrees, certain portions of bone and muſcle, proper for the attainment of light agility, as if aiming at the Mercury's airy form and quality, whilſt on the other hand, ſee the Mercury augmenting his taper figure by equal degrees, and growing towards an Atlas in equal time, by receiving to the like places from whence they came, the very quantities that the other had been caſting off, when, as they approach each other in weight, their forms of courſe may be imagined to grow more and more alike, till at a certain point of time, they meet in juſt ſimilitude; which being an exact medium between the two extremes, we may thence conclude it to be the preciſe form of exact proportion, fitteſt for perfect active ſtrength or graceful movement; ſuch as the Antinous we propoſed to imitate and figure in the mind [2].

I am apprehenſive that this part of my ſcheme, for explaining exact proportion, may not be thought ſo ſum-

[2] The jocky who knows to an ounce what fleſh or bone in a horſe is fitteſt for ſpeed or ſtrength, will as eaſily conceive the like proceſs between the ſtrongeſt dray-horſe and the fleeteſt racer, and ſoon conclude, that the fine war-horſe muſt be the medium between the two extremes.

M 2

Figure 5. Estimation of a line segment using horizontal projection which results in bounding boxes segmenting lines from one another.

etc.) can be inferred through a graph structure, which allows for the flexible categorization of the page and its constituent parts (Fig. 6). The plurality of views of an individual page's layout structure allows researches to label the document according to multiple hypotheses or "representations."[25] For DIA researchers, the page image consists not of a monolithic, singular unit, but rather frames a multiplicity of possible perspectives. It affords a hermeneutic or interpretive relationship to any given page; the very concept of the page image presupposes the semiotic excess of bibliographic objects. In this way, DIA is a direct inheritor of the hermeneutic perspectivalism that has a long history within the field of book history and the assumed polysemy that has remained central to the methods of philology and literary studies since at least the late nineteenth century.

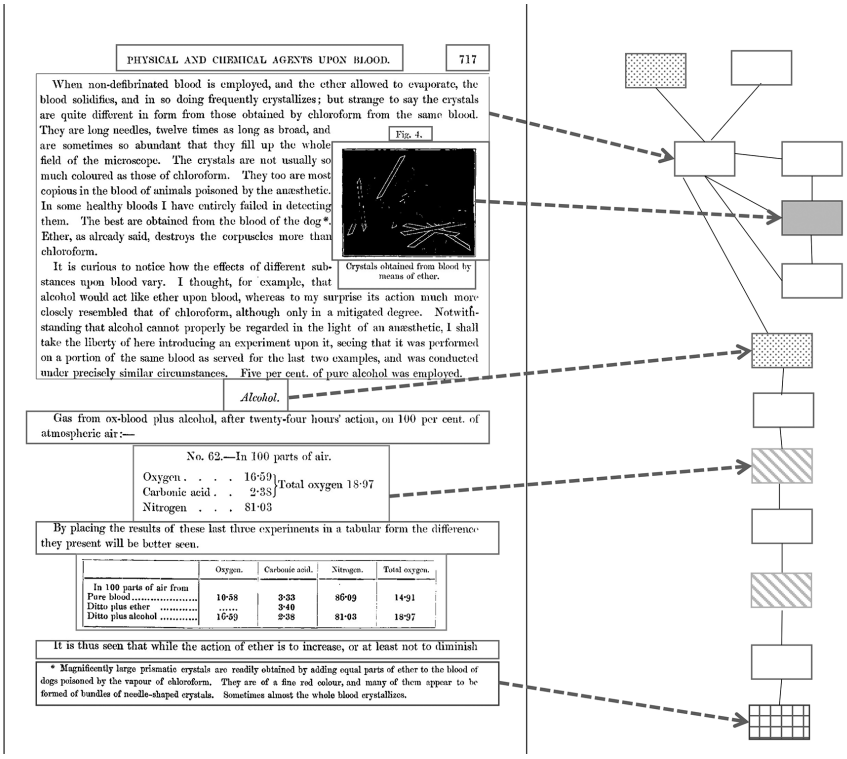


Figure 6. Adjacency graph representation of different page segments. Based on the article by Simone Marinai, "Page Similarity and Classification," *Handbook of Document Image Processing and Recognition.*

Once pages have been segmented, researchers can then focus on identifying and measuring individual *features*. As in any interpretive model, features are potentially limitless. Almost anything on a particular page could be considered a feature. The height of characters or lines, the number of conjoined letters, the width of margins, the presence of illustrative segments—these are all features that can be captured by DIA. Some, like character height (or font size), are more elementary than others such as "visual ornamentation" or "pictorial realism." The most elementary feature is the pixel. Importantly, features are also nested: pixels are used to understand "line height" which might be part of identifying a higher order feature like "title page." The identification of features is a core part of the research process, one that cannot be divorced from specific research questions. How researchers "model" a page determines what they "see." There is no single, universal page view.

As these examples have hopefully made clear, DIA adopts and adapts some of the core analytical values of the field of book history as it has evolved over the course of the past three decades, but it approaches these values less as stable taxonomies and more as contingent frames for interacting with a page. Whether it is the act of "foregrounding," the page "view," or the identification of a "feature space," DIA does not rely on a static ontology of either book or page. Instead, it encodes researchers' beliefs about the social, creative, and epistemological functions of books from different historical epochs or cultural spaces into the process of seeing the page. In this, it is strongly aligned with the interpretive impulses of bibliographical and philological traditions. DIA can foreground specific dimensions of past bibliographic practices alongside current disciplinary ways of understanding them. Computation is not "objective" in the sense that researchers can use computational methods to produce a fixed and "accurate" representation of a page that can then be universalized. In many ways, DIA marks a departure from earlier aims of creating more universal bibliographic taxonomies of the visual dimensions of pages, as in handbooks like Gaskell's. The ornamental focus of Gaskell's system—the attention to typeface, swash letters, rule, and pieces for example—is but one way of thinking about the larger social process of how books' graphic practices participate in generating meaning and organizing social interactions.[26] Current computational approaches to seeing pages can thus be valuable tools in developing and thinking about the broader social history of books.

Table 2.
Procedures for understanding the visual page in DIA.

Category	*Procedure*	*Features*
Normalization	*Binarization*	uneven color-gradient bleeding (through the page, across the page, between characters)
	Deskewing	orientation warping
Analysis	*Segmentation*	characters lines regions
	Feature Identification	pixels characters line height line location line spacing etc.

The Curious Difficulty of Detecting Footnotes: A Case Study

If these are some of the ways that Document Image Analysis draws on book historical and philological methods and practices to construct visual models of the book, in this section we provide a case study of the core analytical steps that are necessary for the *application* of these methods to historical study. How can we use these techniques to learn something about the past? The steps that are essential here belong to the area of machine learning, which entails the process of learning representations of page images in order to generalize about their large-scale prevalence in the world. While there are a variety of technical considerations (which algorithms to use, how to train them, etc.), we focus here largely on the non-computational dimensions of machine learning, once again with the aim of facilitating cross-disciplinary collaboration. We want to foreground, in particular, how the expertise and methodological techniques of book historians become necessary complements to computational methods when these techniques are applied to the study of the past.

For the purposes of this essay, we focus on the case of detecting footnotes, which comprise one of our four primary visual features that we

posit as fundamental to the graphic practices of scientific communication (alongside tables, diagrams, and figures). In our larger project we utilize two principal data sets: the first consists of a collection of proceedings from natural academies of science from five different national contexts (France, Germany, England, Sweden, Russia), dating from between 1665 and 1946 and containing 828 volumes and 512,516 pages. The second consists of the Eighteenth-Century Collections Online database (ECCO), comprising over 32 million pages. Together, these two collections allow us to study the diachronic evolution of the page image across different national contexts in a single scientific genre as well as the synchronic relationships of pages across a more complex discursive environment constrained by a single pivotal historical period and national context.

In order to undertake our analysis, we propose the following four basic steps to implementing the large-scale study of historical documents:

1. Definition
2. Annotation
3. Feature Identification
4. Validation

In order to detect a visual feature (in this case, footnotes) at large scale, we need a working definition of our object of study. For our purposes, we define a footnote in the following way:

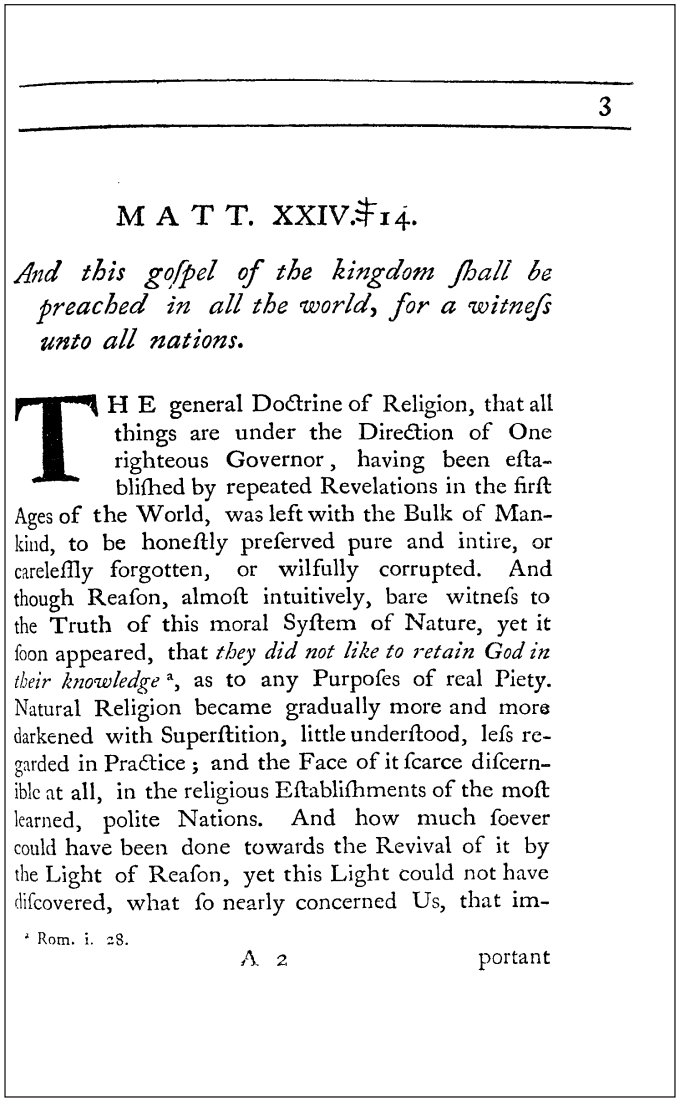
> **Footnotes need to be distinct, marked text at the bottom (foot) of the page that are referenced in the main part of the text.**

However seemingly straightforward such a definition may be, when it comes to generalizing across large collections of historical documents we can find numerous instances of pages that pose problems for our analysis. Consider, for example, this page from *A sermon preached before the Incorporated Society for the Propagation of the Gospel in Foreign Parts* by Joseph Lord Bishop of Bristol (1739) (Fig. 7). Here we have three separate text segments at the bottom of the page. While only one of them (on the left) is "marked" by a footnote mark that matches the body of the text (after the italicized word "knowledge"), one can see how the footnote marks in both the note and body text are extremely small and irregular. Human vision and training allow us to differentiate between a mark that is "noise" (or accidental in our terminology above) and one that is meaningful (or intentional). But think of all the possible confounding associations that this mark could be

interpreted as—given the low resolution of the image how can we associate the mark in the footnote with that in the body? How are we to differentiate between an accidental smudge and this small set of pixels? When it comes to the body of the text, how is the footnote mark different from an apostrophe, imperfection, or quotation mark?

These are just a short list of problems that definition and generalization pose with respect to historical study at large scale. By imagining the amount of learning required for a young person to correctly identify the footnote on this page, one can begin to intuit the challenge a machine will have and the diversity of examples needed to gradually build up an understanding of the concept, even of such a simple idea as a "footnote." Adrian Johns's claims about the invention of print culture as the gradual development of distinct and learned habits of interaction and imagination became an immediate challenge for us in the research process.[27] As we discuss in the next step, we had to transmit to young student research partners assumptions and habits related to printed objects that were, for us, hard won but now largely taken for granted and that we had never had to consciously articulate. Large-scale historical study requires high levels of reflective explicitation: the articulation and clarification of even the most mundane assumptions and behaviors of scholarly practice.[28]

The challenges of definition and the sharedness of definition bring us to the second step in the process: *annotation*. After we have defined our visual feature of interest, we then need to go through as many pages as possible and annotate them for the presence/absence of a footnote such that a machine can eventually learn to properly identify it. As the challenges above indicate, not all cases are straightforward. Decisions will depend on the training of the faculty and students involved in the process as well as the mechanisms used to adjudicate ambiguity. In this step, we manually annotated 22,056 page images for training (6,042 pages with footnotes and 16,014 pages without), and another 5,520 pages for testing (554 with and 4,966 without). Notice how we try to use imbalanced classes of pages (i.e. with footnote and without) to mirror the anticipated prevalence of footnotes in our collections. This is important to avoid what is called overfitting, which will result in poor generalization when applied to a much larger (or diverse) collection of documents. No matter how many documents are annotated in this way, however, it is a tiny fraction of the entire collection. We cannot be certain that our annotations will lead to reliable scaling, which is why we have a necessary final step of the process called *validation*. Important for this step is that the underlying representation that will be learned by the machine is based on a social consensus generated by a contingent



3

MATT. XXIV. ℣ 14.

And this goſpel of the kingdom ſhall be preached in all the world, for a witneſs unto all nations.

THE general Doctrine of Religion, that all things are under the Direction of One righteous Governor, having been eſtabliſhed by repeated Revelations in the firſt Ages of the World, was left with the Bulk of Mankind, to be honeſtly preſerved pure and intire, or careleſſly forgotten, or wilfully corrupted. And though Reaſon, almoſt intuitively, bare witneſs to the Truth of this moral Syſtem of Nature, yet it ſoon appeared, that *they did not like to retain God in their knowledge* [a], as to any Purpoſes of real Piety. Natural Religion became gradually more and more darkened with Superſtition, little underſtood, leſs regarded in Practice; and the Face of it ſcarce diſcernible at all, in the religious Eſtabliſhments of the moſt learned, polite Nations. And how much ſoever could have been done towards the Revival of it by the Light of Reaſon, yet this Light could not have diſcovered, what ſo nearly concerned Us, that im-

[a] Rom. i. 28.

A 2 portant

Figure 7. *A sermon preached before the Incorporated Society for the Propagation of the Gospel in Foreign Parts* by Joseph Lord Bishop of Bristol (1739). Source: Eighteenth-Century Collections Online.

group of researchers. It is the outcome of a social process. The advantage of this is that it mirrors the aims of doing large-scale analysis where the goal is to identify a category that is independent of a single researcher's point of view. The intention behind defining a footnote is to frame the concept

in a way that others would agree with. The practice of annotation tests this assumption. Is this a definition for which we can achieve high levels of interpersonal agreement? Although we did not include this in our process, a key step is to test this degree of consensus through the process of what is known as "inter-rate reliability." This allows researchers to reflect on the degree of sharedness surrounding their definition among different kinds of readers or scholars.

The third step is the most creative: here we try to identify features that will help in the detection of footnotes. Footnotes are seen as a meta-feature composed of nested layers of smaller features. One method we used, for example, measures the relationships between lines and white space as a way of identifying the location and presence of footnotes (Fig. 8). Another uses line heights projected as vertical histograms to capture the possibility that notes will be differently sized than the body of a text (Fig. 9).[29] Finally, another downsamples the page image into a smaller set of pixels to produce a blurred image and then treats this image as a single vector of black/white pixels (Fig. 10).[30] Essential for each of these cases is the fact that there is no neutral page image prior to our construction of it. We are building into our "views" of the page assumptions, beliefs, and hypotheses about how the visual units of interest function for us. The types of footnotes we capture (and don't capture) are entirely dependent on the way we "view" the page and which aspects of footnotes we choose to focus on. Our attention to line-height for example meant that many overlooked footnotes were similar in that they did not change their font size with respect to the body of the text. This is a key way in which the modeling of the features shapes the types of examples your process will detect.

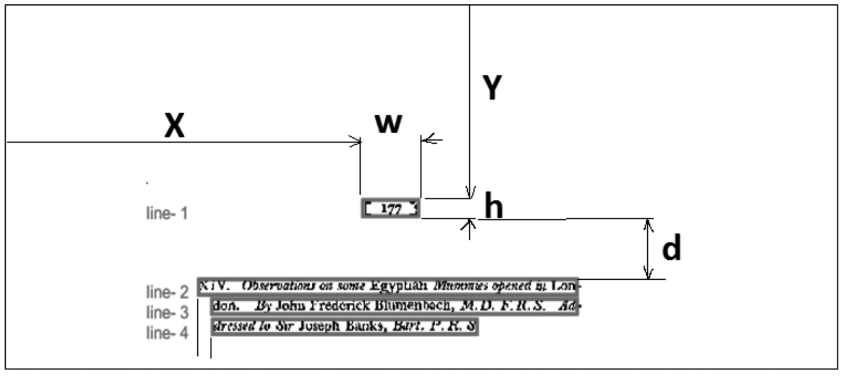


Figure 8. Layout-based features that measure relationships between lines and white space.

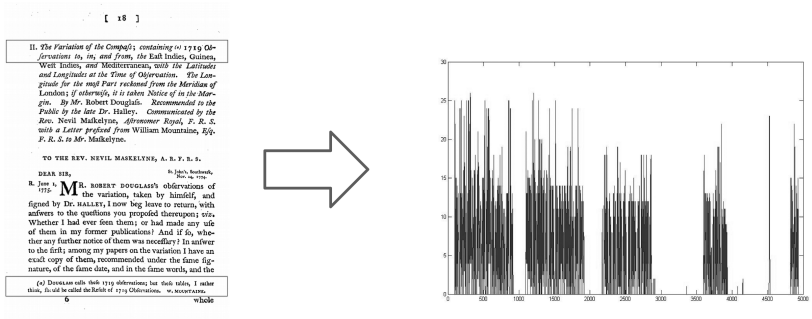


Figure 9. A histogram of line heights for five lines of text from a sample page in ECCO. Here we use vertical projections of the lines, meaning the bars of the histogram represent vertical slices of the red-bounded lines. The lower height of the histogram bars represents a lower average line-height.

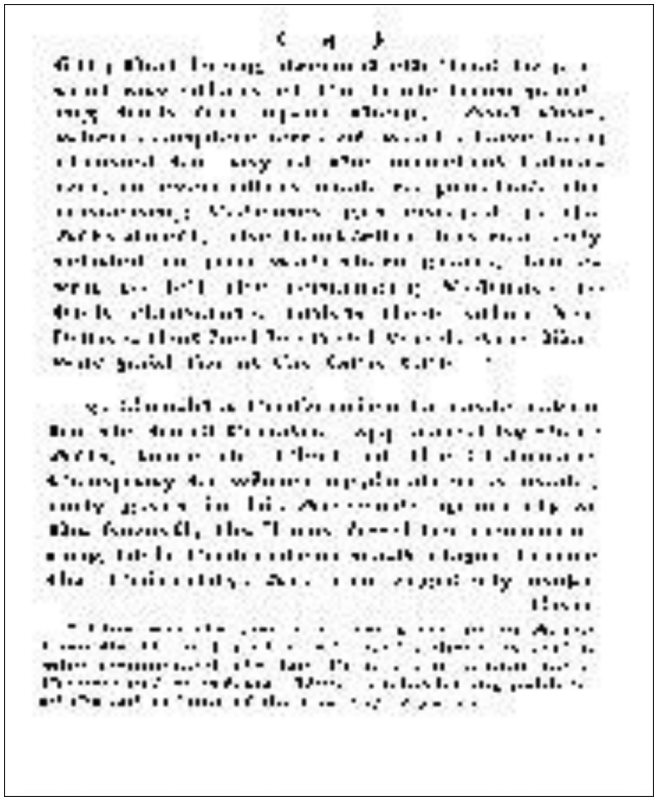


Figure 10. In this example, pages are binarized and then reduced in size to 227x227 pixels (or 51,529 dimensions) rendering them illegible, but ideally capturing the unique visual signature of footnotes.

x = distance from the left edge
y = distance from the top edge
w = width of the textline
h = height of the textline

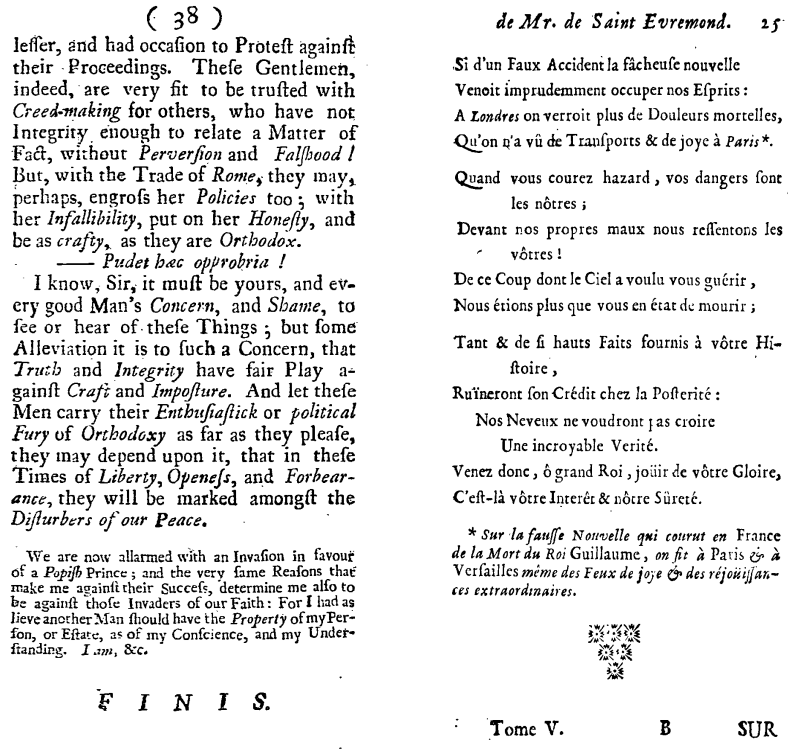
Finally, the last stage is *validation*. Given the use of particular machine-learning algorithms, the selection of which is the provenance of the engineering team, how well do they capture our annotations using these feature representations?[31] Table 3 shows the results of using different algorithms as well as their combined results.

Table 3.
Results of our classification process using four different machine-learning approaches and one ensemble approach.

Approaches	*Precision*	*Recall*	*F1 score*
Rule-based	66.90%	59.80%	0.6315
Layout-based	59.79%	68.52%	0.6385
CNN-based	88.21%	47.40%	0.6166
Transfer learning-based	72.26%	40.49%	0.5189
Ensemble	**94.57%**	**66.98%**	**0.7841**

What this table tells us is that when we identify a page as containing a footnote, using an ensemble of all four algorithms we are "correct" almost 95% of the time. This is the column labeled "precision." However, "recall" measures the extent of pages that were annotated as footnotes that were predicted to be footnotes. Here we see a much lower number. This implies that we only capture 67% of the total number of pages labeled as footnotes in our test set. About one-third of our footnoted pages thus go undetected, but for the ones we capture we are very accurate. This is the classic precision/recall trade-off, for which the most well-known example would be airport security, which relies on the opposite dynamic (very high recall, i.e. lots of people pulled aside for suspicions of carrying a weapon, and low precision, i.e. lots of false positives).

Even more important than the table of values here is a qualitative inspection of the results that necessarily accompanied its production. What kinds of pages are incorrectly identified as having footnotes and what kinds of pages are incorrectly missed that do have footnotes? We include an example of a "false positive" (a page with a predicted footnote that does not have a footnote) and a "false negative" (a page with a footnote that was not predicted to have a footnote) (Fig. 11). As we can see in the page image on the



(38)

leſſer, and had occaſion to Proteſt againſt their Proceedings. Theſe Gentlemen, indeed, are very fit to be truſted with *Creed-making* for others, who have not Integrity enough to relate a Matter of Fact, without *Perverſion* and *Falſhood!* But, with the Trade of *Rome*, they may, perhaps, engroſs her *Policies* too; with her *Infallibility*, put on her *Honeſty*, and be as *crafty*, as they are *Orthodox*.

—— *Pudet hæc opprobria!*

I know, Sir, it muſt be yours, and every good Man's *Concern*, and *Shame*, to ſee or hear of theſe Things; but ſome Alleviation it is to ſuch a Concern, that *Truth* and *Integrity* have fair Play againſt *Craft* and *Impoſture*. And let theſe Men carry their *Enthuſiaſtick* or *political Fury* of *Orthodoxy* as far as they pleaſe, they may depend upon it, that in theſe Times of *Liberty*, *Openeſs*, and *Forbearance*, they will be marked amongſt the *Diſturbers of our Peace*.

We are now allarmed with an Invaſion in favour of a *Popiſh* Prince; and the very ſame Reaſons that make me againſt their Succeſs, determine me alſo to be againſt thoſe Invaders of our Faith: For I had as lieve another Man ſhould have the *Property* of my Perſon, or Eſtate, as of my Conſcience, and my Underſtanding. *I am*, &c.

F I N I S.

de Mr. de Saint Evremond. 25

Si d'un Faux Accident la fâcheuſe nouvelle
Venoit imprudemment occuper nos Eſprits:
A *Londres* on verroit plus de Douleurs mortelles,
Qu'on n'a vû de Tranſports & de joye à *Paris**.

Quand vous courez hazard, vos dangers ſont
les nôtres;
Devant nos propres maux nous reſſentons les
vôtres!
De ce Coup dont le Ciel a voulu vous guérir,
Nous étions plus que vous en état de mourir;

Tant & de ſi hauts Faits fournis à vôtre Hiſtoire,
Ruïneront ſon Crédit chez la Poſterité:
Nos Neveux ne voudront pas croire
Une incroyable Verité.
Venez donc, ô grand Roi, joüir de vôtre Gloire,
C'eſt-là vôtre Interêt & nôtre Süreté.

* *Sur la fauſſe Nouvelle qui courut en* France *de la Mort du Roi* Guillaume, *on fit à* Paris *& à* Verſailles *même des Feux de joye & des réjoüiſſances extraordinaires.*

Tome V. B SUR

Figure 11. Example of a false positive (left), from *An account of the late proceedings of the dissenting ministers at Salters-Hall* (1719). Example of a false negative (right), from *Les veritables oeuvres de Monsieur de Saint-Evremond* (1706).

left, we see a differentiated portion of text that is set off at the bottom of the page and uses smaller font. However, it lacks the footnote mark reference to the body of the text. Because the presence of the footnote mark is so hard to detect, many of the few false positives we predict have this quality. On the right we see how the overall typographic irregularity of the page causes the footnote not to stand out. Further type and ornamentation below the footnote also masks its "footness," i.e. its being located at the bottom of the page. In many other cases, the similarity between the footnote text and the body text is one of the core qualities that appears to allow footnotes to go undetected.

Understanding error is a key component of the research process. It can often help engineer better features for detection. But there are also limits in

how we can generalize about errors. The variability of a practice coupled with the limits of training data (how to identify all possible exceptions to a rule) mean that *uncertainty and error are intrinsically going to be part of the analytical process*. This is a key new dimension to the study of history that requires further reflection as we move forward.

The History of Footnotes

So what can we learn from all of this? Once we are able to detect footnotes at large scale, what new kinds of histories can we tell about print, communication, and knowledge? As we see in Fig. 12, over the course of the eighteenth century in texts printed in the United Kingdom the footnote achieves an interesting normalization in terms of its overall cultural prevalence. Up until the mid-eighteenth century (a statistical model suggests that 1745 is the turning point), the use of footnotes follows a highly predictable linear path of growth. It then levels off into a relatively stable consensus of placing footnotes on around 5.2 percent of all printed pages. This is something we haven't seen before: the way larger cultural behavior follows a rise and then consensus model of production. Whatever we are seeing in terms of local yearly variation, over time there emerges a surprising degree of regularity when it comes to this particular graphic practice.

These findings also open a host of further research questions: Why does the prevalence of footnotes settle at around five percent and what about the fluctuations? Are there semantic markers in titles that are related to increased probabilities of footnoted pages? Is there a discernable relationship between the language of a text and its visual features? We are interested in what these types of questions might tell us about the shape of knowledge in eighteenth-century Europe. These more formal questions quickly lead to questions about the history of knowledge more broadly. What was driving the rise of footnote use in the first half of the century? As Anthony Grafton has argued, the rise of footnotes in the discipline of history ought to be seen as distinct from medieval and early modern practices of bibliographic commentary.[32] How does the increased prevalence of footnotes in the first half of the eighteenth century align with the argument that this was the time period when the "empire of erudition" began to differentiate into more distinct domains of knowledge—that is, when, for example, periodicals devoted themselves not just to reporting the "news" from the Republic of Letters but from increasingly distinct domains of scholarly communication?[33] How

might a better understanding of footnotes (and tables, diagrams, figures, and other graphic features) lead to a better understanding of the emergence of disciplinary knowledge? We don't have definitive answers to these questions yet, but the visual data can help give us interpretive frameworks through which we can better understand these historical communicative practices at large scale.

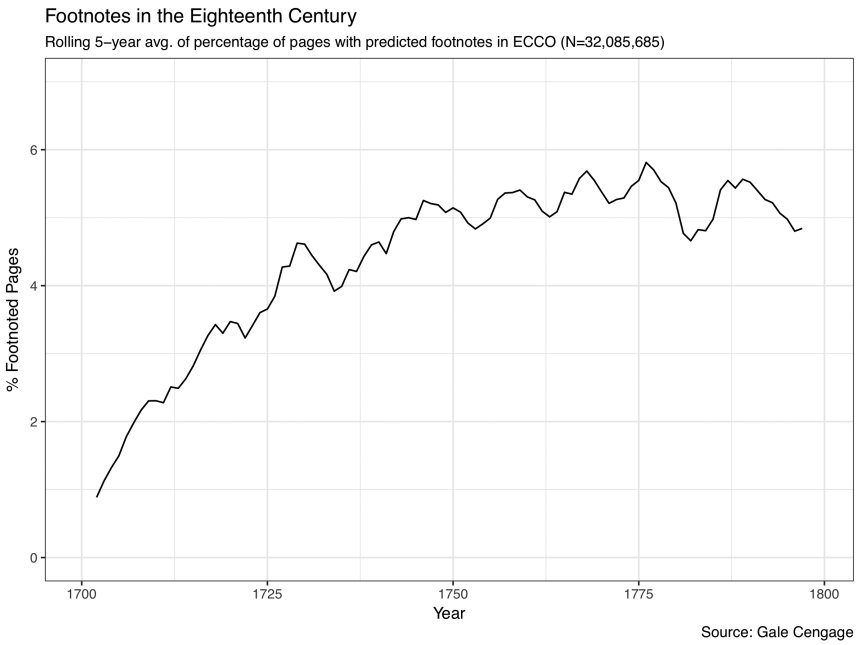


Figure 12. Predicted average number of footnoted pages per year in the Eighteenth-Century Collections Online database.

We can also burrow further into the data to see whether different genres or domains of writing behave differently with respect to footnotes. While somewhat hard to visualize distinctly in black and white, what we see happening is the way history writing is a clear leader in the use of footnotes over the course of the century (Fig. 13). Until the beginning of the nineteenth century, as scholars like Grafton have long argued, historical writing is a much stronger driver of inter-bibliographic citation than scientific writing (the third lowest of our four genres).[34] We also see how the arts begin to distinguish themselves by notably decreasing their use of citation. This trend is particularly notable given that one of the linguistic indicators of books that

are more likely to use footnotes in the first half of the century are related to "belles lettres," rhetoric, and classical philology.[35] The eighteenth century serves as a fascinating period of ideological transformation with respect to creative writing—from a driver of inter-bibliographic citations to one that is increasingly defined by their absence. Reading literature shifts from a practice that is highly mediated by scholarly perspective to one that is designed to be immersive and immediate, but according to our findings it does so at a far earlier time-frame than has traditionally been identified.

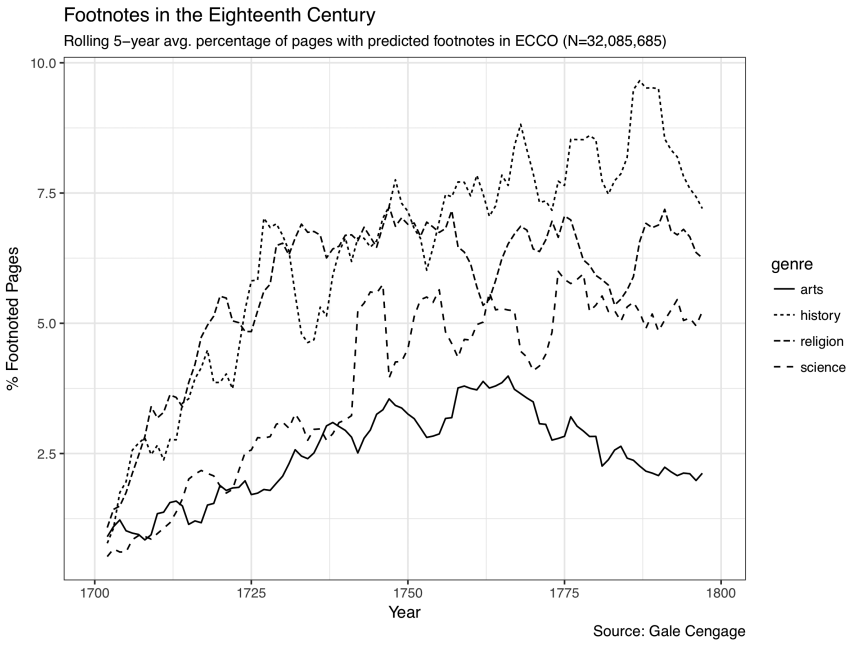


Figure 13. Predicted levels of footnoted pages in the eighteenth century broken out by genre.

Finally, when we look at our other data set of the proceedings of five different European national academies of science (Fig. 14), we see how after 1800 there is a second large wave of citational practices that comes to define scientific writing. The consensus that we saw across genres in the eighteenth century is far surpassed by the particular domain of scientific communication. Here the critical period appears to be the first half of the nineteenth century, in particular 1825–50. While we don't have data on historical writing for this period, we suspect that scientific citational practices

would far outweigh those of history, suggesting a new period of disciplinary configuration. What, if any, is the relationship between the sharp spike in footnoted pages in academy proceedings and, as the historian of science Alex Csiszar has recently put it, the "intrusion of journals into elite scientific institutions" in the 1830s?[36] Did the decision of the academies to publish journals, as opposed to the "weighty" and not exactly periodical tomes like the *Philosophical Transactions*, also invite different printing and reading practices and ways of establishing epistemic authority? These are just some of the preliminary insights and questions that the visual history of print and scientific communication can offer, though a fuller account awaits.

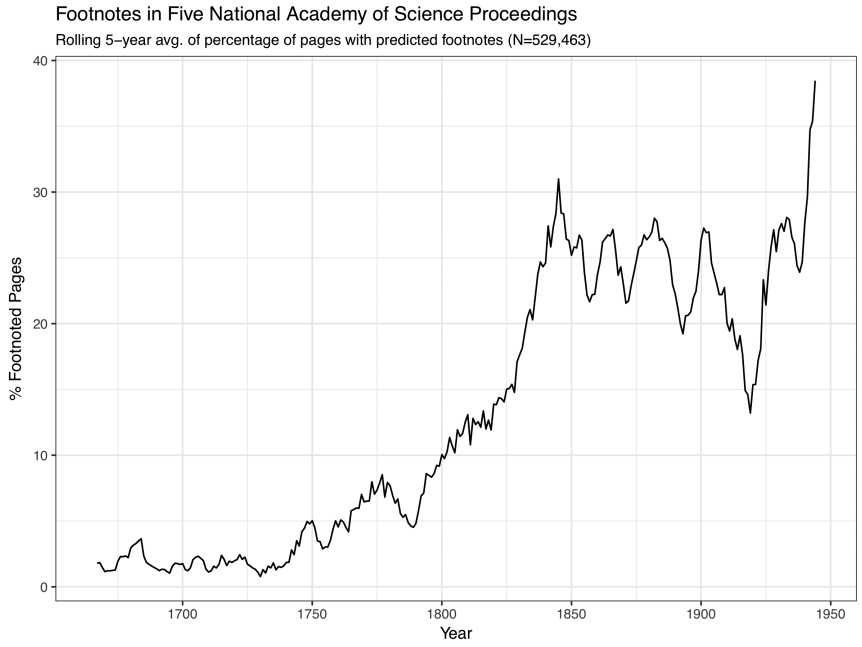


Figure 14. Predicted percentage of footnoted pages in five different proceedings of national academy of sciences.

Conclusion

In this essay, we have tried to explain the computational and epistemological conditions through which page images can become objects of historical study, along with some preliminary insights about what that knowledge might yield for historians. Rather than focus exclusively on individual find-

ings, our primary aim has been to clarify the conditions through which such findings can be made. One of our goals is to expose our fellow book historians to a relatively new method and technique, DIA, for understanding documents and to introduce computer scientists and DIA researchers to well-established scholarly practices and norms long devoted to the study of historical texts. As we have outlined above, although there are significant differences in the ways DIA and book history approach texts, there are also important continuities. As we have shown, DIA, oftentimes implicitly, incorporates basic book-historical and bibliographical insights into its various processes. It has inherited assumptions, categories, insights, and ideals that a simplistic opposition between computational and humanist methods or scholarship obscures. In conclusion, we would like to consider how the continuities and differences between DIA and book history and bibliography we have highlighted can bring into focus some core methodological questions concerning the study of historical documents under the changing conditions of our archives.[37]

1. Procedural knowledge

DIA is a highly procedural form of knowledge. It involves a series of analytical *steps* that are most often taken in a linear fashion. Metaphors of work*flows* or *pipelines* abound in computational research. All of the steps we have outlined here—normalization, segmentation, feature selection, training, prediction, and validation—are carried out sequentially and lead to a synoptic understanding of visual features of large numbers of historical documents. While this may appear to depart from past archival practices of scholars, practices which to outsiders (and even to insiders at times) may have the look and feel of being more haphazard in their discovery process, the manual searching of archives is similarly organized by a theoretical framework that governs the pathways and choices made to navigate a documentary repository that cannot be grasped in its entirety. The difference of DIA, we would argue, is that these pathways are (or ought to be) made explicit at each step. Unlike the scholar in the archive, the behavior of the DIA research team is intended to be fully reproducible. Proceduralism in this sense can be a powerful form of visibility.

2. *Transformational knowledge*

DIA is highly transformational in nature.[38] From binarization to feature selection to machine learning, at every stage we are producing new representations of the underlying page images in order to better understand them at large scale. In one sense, such work is well aligned with the long tradition of editorial labor of reproducing historical documents, where careful and thoughtful engagement with the practices of transformation is essential to the field.[39] On the other hand, such transformational principles also mark a strong departure from the invocations of textual particularity used by McKenzie and his followers, where the goal of such research is the recovery of original documents in as much historical detail as possible. And yet in both cases the ultimate aim is greater knowledge about past *social* practices. The moment of generalization about the past is the moment at which a new representation is created of that past and the original sources are synthesized. As with the construction of critical editions that account for the various witnesses that contribute to a final manifestation, DIA foregrounds the use of mediation to understand the processes of historical mediation.

3. *Contingent knowledge*

Much of the enthusiastic embrace of "big data" over the past decade by scholars and broad swaths of the public has largely been driven by the explosion of relatively well-structured data produced by social media or related internet-based platforms. Our ability to study the past, however, is contingent upon what materials are available and the forms, media, and organization that structure how scholars can engage them. Like all computational methods, DIA depends upon what documents have been digitized and how well they have been organized. Historians and bibliographers take these limitations and contingencies as given conditions of trying to study the past, and so they have developed practices and methods for dealing with working with the contingencies of archival and textual materials. The computational study of the past and the recourse to large scale do not mean we can leave behind these considerations of inclusivity and representativeness. If anything, these questions become even more urgent as we attempt to generalize about past practices based on datasets whose sheer size can all too easily blind us to the exclusions, gaps, and omissions within them.

4. *Knowledge and error*

Bibliographers, philologists, and historians have always thought about "error," whether in terms of textual "corruptions," (sources that misrepresent an underlying ideal expression) or in terms of false inferences about the past on the part of another commentator. But whereas scholars working with texts have traditionally sought to excise or correct such "errors," computational methods such as DIA make error a constitutive feature of the knowledge they produce. This is a key difference. As a conditional form of knowledge, DIA makes the degree of uncertainty part of the inferential process. In producing what amounts to a model of a particular page, DIA provides an account of what it is attempting to measure as well as the limitations of its success in doing so. Rather than only produce categorical judgments about the presence/absence of footnotes (e.g. "Footnote" or "No Footnote"), DIA also provides probabilistic estimates (e.g. the algorithm is 68 percent or 92 percent certain there is a footnote on the page). How those probabilities relate to interpretive judgments about past documents opens up yet another new avenue of research. What kinds of histories of inferential uncertainty can we begin to tell with respect to different technologies, archives, and algorithms? How does the level of uncertainty present in a process impact the kind of historical narratives we might construct? What is the relationship between error and argument when it comes to computational modeling?

As we have tried to show, DIA is a valuable tool for historical research. It can help discover historical phenomena at scales that have simply not been possible with the manual sorting of large documentary archives. In drawing on many of the principles and beliefs of book historians and bibliographers, it inserts important forms of evidence into the computational study of the past. At the same time, DIA also introduces challenging new epistemological conditions for historical interpretation. What are we to make of the multiple transformations that documentary evidence undergoes in the process of machine learning and how should this influence our understanding of the past? How can we incorporate the multiple and fundamentally contingent points of view that guide processes of annotation and validation when it comes to data? Finally, how can we build historical narratives around concepts like uncertainty and error rather than as replacements of these same concepts? These are just some of the questions that DIA introduces to historical research that we will want to grapple with as a community moving forward.

Notes

1. See the *Proceedings of the First International Conference on Document Analysis and Recognition* (ICDAR 1991). In a not surprising twist in the vagaries of document digitization, the proceedings of the first conference on document image analysis were themselves never digitized and subject to analysis.

2. George Nagy, "Twenty Years of Document Image Analysis in PAMI," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, no. 1 (2000): 38–62.

3. Ryan Cordell, "Q i-jtb the Raven: Taking Dirty OCR Seriously," *Book History* 20 (2017): 188–225.

4. Matthew Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge: MIT Press, 2008).

5. Andrew Piper, "Deleafing: The History and Future of Losing Print," *Gramma*, Special Issue on "The History and Future of the Nineteenth-Century Book," edited by Maria Schoina and Andrew Stauffer, Vol. 21 (2013).

6. Chirstoph Windgätter, *Medienwechsel. Vom Nutzen und Nachteil der Sprache für die Schrift* (Berlin: Kadmos, 2006). See also Carlos Spoerhase, *Linie, Fläche, Raum. Die Drei Dimensionen des Buches in der Diskussion der Gegenwart und der Moderne* (Göttingen: Wallstein, 2016).

7. Garrett Stewart, *The Look of Reading: Book, Painting, Text* (Chicago: University of Chicago Press, 2006).

8. Sybille Krämer, ed. *Schriftbildlichkeit: Wahrnehmbarkeit, Materialtät and Operativität von Notationen* (Berlin: Akademie Verlag, 2012); Sybille Krämer, "Zwischen Anschauung und Denken: Zur Epistemologischen Bedeutung des Graphismus," in *Was Sich Nicht Sagen Lässt,* edited by Joachim Bromand and Guido Kreis (Berlin: Akadamie Verlag, 2010). Krämer's work is an attempt to rethink classical accounts of writing that are tied to orality and alphabetization. See Walter Ong, *Orality and Literacy: The Technologizing of the Word* (London; New York: Methuen, 1982).

9. See Mara Mills, "Print Disability. Die Ko-Konstruktion von Blindheit und Lesen," *Disability Trouble*, ed. Ulrike Bergermann (Berlin: b_books, 2013) 195–204.

10. See Mark Changizi, *The Vision Revolution: How the Latest Research Overturns Everything We Thought We Knew About Human Vision* (Dallas: BenBella Books 2009), and Stanislas Dehaene, *Reading in the Brain: The New Science of How We Read* (New York: Penguin, 2009).

11. Paul Fyfe and Qian Ge, "Image Analytics and the Nineteenth-Century Illustrated Newspaper," *Journal of Cultural Analytics,* October 24, 2018, DOI: 10.22148/16.026, and Neal Audenaert and Natalie M. Houston, "VisualPage: Towards Large Scale Analysis of Nineteenth-Century Print Culture," *2013 IEEE International Conference on Big Data* (2013), 9–16.

12. Elizabeth Eisenstein, *The Printing Revolution in Early Modern Europe* (Cambridge: Cambridge University Press, 1983); Adrian Johns, *The Nature of the Book: Print and Knowledge in the Making.* (Chicago: University of Chicago Press, 2000); Ann Blair, *Too Much to Know: Managing Scholarly Information Before the Modern Age* (New Haven: Yale University Press, 2011); Aileen Fyfe and Bernard Lightman, eds., *Science in the Marketplace: Nineteenth-Century Sites and Experiences* (Chicago: University of Chicago Press, 2007).

13. Bradley Pasanek and Chad Wellmon, "The Enlightenment Index," *Eighteenth-Century Theory and Interpreation* 56, no. 3 (2015): 257–80; Anthony Grafton, *The Footnote: A Curious History* (Harvard: Harvard University Press, 1999).

14. John Bender and Michael Marrinan, *The Culture of the Diagram* (Palo Alto: Stanford University Press, 2010).

15. Lorraine Daston and Peter Galison, *Objectivity* (Cambridge: Zone, 2007), and Barbara Maria Stafford, *Artful Science: Enlightenment and Entertainment and the Eclipse of Visual Education* (Cambridge: MIT Press, 1996).

16. Lorraine Daston and Peter Galison, *Objectivity* (Cambridge: Zone, 2007).

17. See two recent special issues devoted to the topic in literary studies: Krishan Kumar and Herbert F. Tucker, eds., "Special Issue. Writ Large," *New Literary History* 48, no. 4 (2017); James English and Ted Underwood, eds., "Special Issue. Scale and Value: New and Digital Approaches to Literary History," *MLQ* 77, no. 3 (2016).

18. Philip Gaskell, *A New Introduction to Bibliography* (New Castle: Oak Knoll, 1995) 154–59.

19. Theodor Kutschmann, *Geschichte der deutschen Illustration, vom ersten Auftreten des Formschnittes bis auf die Gegenwart*, 2 vols. (Berlin: F. Jäger, 1900); Arthur Rümann, *Das Illustrierte Buch des XIX. Jahrhunderts in England, Frankreich und Deutschland, 1790–1860* (Leipzig 1930); Gordon N. Ray, *The Illustrator and the Book in England from 1790 to 1914* (Oxford: Oxford University Press, 1976); and John Harthan, *The History of the Illustrated Book: The Western Tradition* (London: Thames and Hudson, 1981).

20. See Andrew Piper, "Adapting," *Dreaming in Books: The Making of the Bibliographic Imagination in the Romantic Age* (Chicago: University of Chicago Press, 2009) 183–234; Andrew Piper, "Vanishing Points: The Heterotopia of the Romantic Book," *European Romantic Review* 23, no. 3 (2012): 381–91; and The Multigraph Collective, *Interacting with Print: Elements of Reading in the Era of Print Saturation* (Chicago: University of Chicago Press, 2018).

21. Our assessment of the practices of DIA are based on the following two handbooks, which we strongly recommend to readers interested in this field: Mohamed Cheriet et al., eds., *Character Recognition Systems* (Hoboken: Wiley, 2007) and David Doerman and Karl Tombre, eds., *Handbook of Document Image Processing and Recognition* (London: Springer, 2014).

22. See D.F. McKenzie, *Bibliography and the Sociology of Texts* (Chicago: University of Chicago Press, 1999), and more recently the special issue of *Memoires du livre* on "Textual Histories," where these principles are put into practice. Yuri Cowan, ed., "Textual Histories," *Memoire du livre* 4, no. 2 (2013).

23. Cordell, "Q i-jtb the Raven."

24. Andreas Dengel and Faisal Shafait, "Analysis of the Logical Layout of Documents," in *Handbook of Document Image Processing and Recognition*, edited by David Doermann and Karl Tombre (London: Springer, 2014), 188.

25. Dengel and Shafait, "Analysis of the Logical Layout of Documents," 188.

26. See The Multigraph Collective, *Interacting with Print*.

27. Johns, *The Nature of the Book*.

28. This is one of the values of large-scale historical study. Far from leaving behind close, immediate understandings of individual objects or documents, successful scaling-up depends first on an intense level of scaling down to understand an object of study in its most particular form.

29. Mohamed Mhiri, Sherif Abuelwafa, Christian Desrosiers, and Mohamed Cheriet, "Footnote-based Document Image Classification using 1D Convolutional Neural Networks and Histograms," *International Conference on Image Processing Theory, Tools, and Applications* (2017).

30. Sherif Abuelwafa, Mohamed Mhiri, Rachid Hedjam, Sara Zhalehpour, Andrew Piper, Chad Wellmon, and Mohamed Cheriet, "Feature Learning for Footnote-Based Document Image Classification," *International Conference on Image Analysis and Recognition* (2017): 643–50, DOI: 10.1007/978–3–319–59876–5_71.

31. For the purposes of this essay we omit lengthier discussions of the algorithms used in the project and how that impacts the process. For a fuller discussion, see *Sherif Abuelwafa,*

Sara Zhalepour, Ehsan Arabnejad, Mohamed Mhiri, Emilienne Greenfield, James P. Ascher, Sofia Bach, Victoria Svaikovsky, Alayne Moody, Andrew Piper, Chad Wellmon, and Mohamed Cheriet, "Detecting Footnotes in 32 Million Pages of ECCO," Journal of Cultural Analytics, December 3, 2018, DOI: 10.31235/osf.io/7m8ue.

32. Grafton, *The Footnote*.

33. Chad Wellmon, *Organizing Enlightenment: Information Overload and the Invention of the Modern Research University* (Baltimore: Johns Hopkins University Press, 2015), 45–76.

34. Grafton, *The Footnote*.

35. We use an information-theoretic measure of a g-test to identify words that are more likely to appear in titles of highly footnoted books than titles at large during the transitional period 1700–1745.

36. Alex Csiszar, *The Scientific Journal: Authorship and the Politics of Knowledge in the Nineteenth Century* (Chicago: University of Chicago Press, 2018), 6.

37. For a helpful discussion of the challenges confronting those hoping to use data-driven and computational methods in cultural sociology, see Christopher A. Bail, "The Cultural Environment: Measuring Culture with Big Data," *Theoretical Sociology* (2014) 43: 465–82.

38. For a discussion of the transformational nature of algorithms, see Benjamin M. Schmidt, "Do Digital Humanists Need to Understand Algorithms?," *Debates in Digital Humanities* 48 (2016).

39. See Darcy Cullen, ed., *Editors, Scholars, and the Social Text. Studies in Book and Print Culture* (Toronto: University of Toronto Press, 2012).