Meaning and mining: the impact of implicit assumptions in data mining for the humanities

D. Sculley
Tufts University, Somerville, MA, USA

Bradley M. Pasanek
University of Virginia, Charlottesville, VA, USA

Abstract

As the use of data mining and machine learning methods in the humanities becomes more common, it will be increasingly important to examine implicit biases, assumptions, and limitations these methods bring with them. This article makes explicit some of the foundational assumptions of machine learning methods, and presents a series of experiments as a case study and object lesson in the potential pitfalls in the use of data mining methods for hypothesis testing in literary scholarship. The worst dangers may lie in the humanist's ability to interpret nearly any result, projecting his or her own biases into the outcome of an experiment—perhaps all the more unwittingly due to the superficial objectivity of computational methods. We argue that in the digital humanities, the standards for the initial production of evidence should be even more rigorous than in the empirical sciences because of the subjective nature of the work that follows. Thus, we conclude with a discussion of recommended best practices for making results from data mining in the humanities domain as meaningful as possible. These include methods for keeping the the boundary between computational results and subsequent interpretation as clearly delineated as possible.

Correspondence:
Brad Pasanek, English Department, University of Virginia, PO Box 400121, 427 Bryan Hall, Charlottesville, VA 22904-4121, USA.
E-mail:
bmp7e@virginia.edu

A person who is trying to understand a text is always projecting. He projects a meaning for the text as a whole as some initial meaning emerges in the text. Again, the initial meaning only emerges because he is reading the text with particular expectations in regard to a certain meaning. Working out this fore-projection, which is constantly revised in terms of what emerges as he penetrates into the meaning, is understanding what is there.

– Hans Georg Gadamer, *Truth and Method*

1 Introduction

In any new relationship—scholarly or otherwise—it is what is *not* said that can be most harmful. Coming from disparate backgrounds, we tend to believe that our own assumptions are shared assumptions. When these are left tacit at the outset, they may become all the more damaging when it is discovered that beliefs thought to be self-evident by one party are considered surprising or outlandish by the other. The collaborative field of

Literary and Linguistic Computing, Vol. 23, No. 4, 2008.
doi:10.1093/llc/fqn019 Advance Access Published on 12 September 2008

digital humanities, still in its nascency, is characterized by new collaborations between scholars of language and literature on the one hand, and computer scientists on the other. It is unlikely that the assumptions of one field are held as tenets of the other. Furthermore, because these fields are so different, it may be difficult for scholars in each field to articulate and pose questions for one another. We do not know what it is that we do not know, or even what to ask of each other.

In this article, we attempt to bridge the disciplinary divide by making some of these assumptions explicit. In particular, we examine literary methods and literary critical premises, placing them in contrast with the implicit assumptions underlying data mining and machine learning methodologies. In order to establish a ground for this conversation between the disciplines, we work our way through a case study, employing data mining techniques in hypothesis testing. Specifically, we investigate a purported connection between metaphor and political affiliation, using a classifier to predict an author's party politics. Needless to say, this venture remains fraught with peril. As will be apparent, it is all too easy to draw mistaken conclusions when using the powerful tools afforded by machine learning.

2 Background: Fears of Circularity

In his masterpiece, *Truth and Method*, Hans-Georg Gadamer presents a philosophy and history of interpretation, following interpretive practices from their scriptural origins through Schleiermacher, Dilthey, and Heidegger. Gadamer writes, 'A person who is trying to understand a text is always projecting . . . working out [a] fore-projection, which is constantly revised in terms of what emerges' (p. 267). The interpreter is implicated in circularity because he must project meaning onto the very object that he hopes will disclose its meaning for him. After selecting literary material and selecting features of that material for analysis, we hope that meaningful patterns emerge from out of what is already there. But we must always take into account the powerful determinations of our own fore-projections.

The digital humanities, still struggling to establish its legitimacy, is characterized by its own peculiar anxieties about circularity. How do the results of computer-assisted textual analysis (often packaged as graphs and scatter plots) come to speak to a traditional literary critic? The critic in question would have read the plays, poems, or novels that serve as the basis of the mining, clustering, or classifying experiment, and he or she would either come to find his or her previous readings of the texts confirmed in the abstract model or dismiss the results and the validity of the computational methods that produced them. The model itself must be interpreted, and the critic brings his prior readings to bear on the abstract representation in just the same way he would bring them to bear on the text itself. The critic reads varieties of confirmation into the results or dismisses them.[1]

What can the computer tell the literary scholar—that know-it-all who is ever ready with an interpretation, who can assimilate almost any piece of information to his or her pet theory? At *The Valve* Matthew Kirschenbaum reports the results of data mining experiments with Emily Dickinson's correspondence. One of the Dickinson scholars involved in the experiments saw that the word 'mine' ranked high on the list of words that had been predicted as indicators of erotic language. She reflects, 'The minute I saw it, I had one of those ''I knew that'' moments.' The scholar finds occasion to stage a reading of the term 'mine'[2] and delivers the following appraisal of text mining methods: 'So the data mining has made me plumb much more deeply into little four- and five-letter words, the function of which I thought I was already sure, and has enabled me to expand and deepen some critical connections I've been making for the last 20 years.' Although the scholar registers surprise ('I thought I was already sure'), we hear her own interpretative machinery assimilating the computer-generated results with alarming ease. Who has *plumbed* more deeply here? The machine or the scholar? And while the 'critical connections' are expanded and deepened, they are not revised, replaced, or overthrown.[3]

The interpreter of a poem moves from the words, sentences, verses, and stanzas to the largest structure to which they contribute. The text is a whole, and to

understand it is to understand its parts.[4] But the text itself is situated in a larger context of texts, a whole tradition of texts. In the recent quantitative turn taken in Franco Moretti's scholarship, we find an exciting polemic against the production of 'readings'. Moretti would have us direct our efforts at explanation, not interpretation and demands a complete account of the novel or, even more grandly, the world system of literature as the goal of literary history. His methods do not offer 'a new reading' of a canonical novel, such as *Waverley*[5] 'but aim to understand the larger structures within which [a novel like *Waverley* comes to] have meaning in the first place' (p. 63). Moretti would have us change our basic unit of analysis: instead of the individual text we should look to the great unread multitude of texts that compose a genre. We find Moretti's mention of *Waverley* particularly appropriate. *Waverley* is, after all, the novel that Georg Lukács treats as representative of historical fiction as such. Scott's 'average' or 'typical' hero, Waverley, stands in the center of a greater socio-political history. The novel's plot seeks a 'neutral ground' in which we find 'extreme, opposing social forces . . . brought into human relationship with one another' (Lukács, 1983, p. 36). Reading *Waverley* we make it stand for something larger; we read it, in fact, under Lukács' influence, as itself involved in those larger structures we look to in order to gauge a particular text's meaning 'in the first place'. Again, we cannot help but trace a hermeneutic circle, moving from a textual instance to some near totality and back.

Moretti's promotion of 'distant reading' practices gains ground in a digital humanities, in large part, because the practices are well suited to the projects imagined by those who seek new research opportunities in an online digital 'million book' library. It is predicted that in working with the largest wholes, we digital humanists will be in a position to make better sense of the relations that govern the innumerable parts. While questions remain as to how to work with so much material, there is a growing consensus that machine learning may allow us function at this highest level of granularity. The leading edge of 'computation into criticism' may then be moving away from text analysis to data mining and other machine learning methods. Speaking about collaborative efforts, John Unsworth has claimed 'The computer provides us with the ability to keep track of enormous amounts of information, to sort and select that information rapidly and in many different ways, and to uncover in reams of mute data the aesthetically and intellectually apprehensible patterns on which understanding depends' (Unsworth, 1997).

3 Assumptions in Data Mining

Is distant reading amenable to machine learning? The tools of data mining and machine learning are built on the foundation of computational learning theory. First formalized in Valiant's 1984 paper 'A Theory of the Learnable' (Valiant, 1984), learning theory provides guarantees on the learnability of arbitrary concepts, based on certain fundamental assumptions. These assumptions are at odds with those of the average humanities scholar—thus, it is worth stating them explicitly here.

3.1 Examples are drawn from a fixed distribution

Learning theory assumes that the data is produced by some process with constant probabilistic qualities. Examples are drawn identically and independently (*i.i.d.*) from this distribution. The key is that the distribution's probabilistic behavior *does not change over time*, and that it will continue to produce as many examples as requested. This assumption enables a formalization of the idea of generalization, and allows bounds to be proved demonstrating that a hypothesis learned on previous data may be able to predict labels for future data.

In data mining for the humanities, this assumption rarely if ever holds. Literary critics almost always work with historical data and attend to some important moment of change: an author's formation of a 'late style', the establishment of vernacular literature in the shadow of a classical inheritance, the refinement and reinvention of the sonnet cycle in the early modern period, or the rise or elevation of the novel in the eighteenth century to name a few. Furthermore, the data sets they work

with are often fixed or limited in size. Although it is common practice in data mining to use fixed training and test sets to approximate the effect of an *i.i.d.*, we must be wary when the size of the data sets becomes too restricted. Typically, English literature monographs explore a thesis in the selected works of five or six authors. Thus, we must pay strict attention to the manner in which the data sets are constructed, as biased sampling (in which the *i.i.d* assumption is broken) may radically skew the qualities of the approximated distribution.

Perhaps even more troubling for many humanities scholars, the idea of a fixed distribution implies that there is, at some level, a fixed truth. Data mining tools *require* this assumption of a fixed distribution to be true, yet the mere mention of 'truth' may make some literary critics uneasy. Literary criticism has spent generations exploring textual ambiguity and instability—reveling in 'the pleasure of the text'. While this may be the single core principle of much literary theory (coded variously as 'ambiguity' or '*différance*'), it is paradoxical in many of its formulations at best. Pleasure is subjective. *De gustibus non disputandem est*. Interpretations are relative to a theory, to a time, or to a reader—to a palate. Outside of a narrowly circumscribed interpretive community, there is no best way to interpret a text. In recent years especially, literary critics have relied on a kind of theoretical promiscuousness or *bricolage* when approaching texts, borrowing from various theories and theorists as necessary to undergird or emphasize a point.

3.2 The hypothesis space *C* is restricted

A *hypothesis space* is the set of all possible concepts that the learning method might produce. Learning theory requires that this hypothesis space be subject to defined restrictions, otherwise generalization cannot be guaranteed.

To see this intuitively, note that with no restrictions a learner can always 'learn' on training data with 100% accuracy simply by rote memorization of the training data. Yet, this rote learner may have no ability to predict a label of a new, unseen example.

The experimenter must therefore decide what form of hypothesis is most appropriate, and this necessarily influences the choice of learning algorithm used for the data. Are we searching for a linear decision boundary in high-dimensional space? Such a representation would require a learning method such as logistic regression, support vector machines (SVMs), or variants of the perceptron or winnow algorithms. Do we believe there is a clearly observable statistical process producing the data points? In these cases, a generative model such as a naive Bayesian classifier, Bayesian networks, hidden Markov models would be most appropriate. Perhaps the hypothesis would be best represented as a set of logical conjunctions, suggesting decision trees. Here, again, experimenter assumptions operate at the forefront and directly impact the design of experiment and subsequent results. And it is hard to say a priori what the best hypothesis space is for a particular literary exploration, such as distinguishing metaphor from metonym or isolating a brute fact of the literary marketplace.

3.3 The data is well represented

The first two assumptions are expressed in terms of abstract distributions and hypothesis spaces. However, for results on real data to be meaningful, the *representation* of the data must be sufficient to capture the salient qualities of that data. Because these representations are, necessarily, simplified abstractions of the real objects, there is a danger that these simplifications are inadequate for capturing important qualities. Worse still is the possibility that such a projection may introduce distortions or procedural artifacts into our representation of the data.

For example, consider the problem of representing a piece of text. One common methodology is to use the 'bag of words' model, where a document is represented by a histogram, with a unique bin for each word that contains a score related to the number of times that word appears in the document. This representation has been extremely successful in the text classification domain (Mitchell, 1997), despite the obvious drawback that all sequence information of the words is discarded. If we decide to employ a bag of words model, the selection of a term scoring method requires an additional decision that impacts the representation.

In information-retrieval tasks, TF–IDF (term frequency–inverse document frequency) scores have been successful (Salton and Buckley, 1988); in the spam filtering domain, binary term weights have proven superior (Metsis *et al.*, 2006). Other possibilities include giving terms higher scores if they appear closer to the beginning of the document, or using some form of semantic analysis to propagate term scores to related terms in the histogram.

Moreover, it is not clear that the bag of words model is itself always optimal. In some cases, counting phrases may be more important; in other cases, tagging parts of speech or tracking sequences of words is paramount. In still other cases, the optical characteristics of the actual letters are of greatest import. The simplest poem is an interaction of grammar, assonance, consonance, enjambment, trope, allusion, and meter. Complexity is multiplied when we work with discourse at large and try to represent a motley collection of poems, novels, essays, and plays. Consider now the difficulty of designing the most appropriate feature representation for impressionist description or political rhetoric. In every case, the experimenter must make assumptions about what is most important in the data.

3.4 There is no free lunch

At first, it may appear that the answer to these problems of experimenter bias in feature representation and hypothesis space is to create a *meta-learner*, that searches through all the different hypothesis spaces, and all the different feature representations to find the optimal design. However, this is provably impossible, as demonstrated by the No Free Lunch Theorem (Wolpert and Macready, 1995).

In traditional data mining applications, the No Free Lunch Theorem means that there is no single best learning algorithm, and we may have to employ a good deal of ingenuity to learn from difficult data. In data mining for the humanities, where learning algorithms are not necessarily used for prediction but instead for exploration and hypothesis testing, the implications are somewhat more severe. In these latter applications, data mining methods will always be subject to experimenter bias. While this does not mean that data mining methods are not a useful set of pattern-finding tools for humanities scholars, we must take particular care to explicate the exact set of implicit assumptions we make before drawing interpretations from any results. Some of these challenges and potential pitfalls are illustrated in the following section.

4 An Experimental Case Study

> *The first principle is that you must not fool yourself – and you are the easiest person to fool.*
> – Richard Feynman, 'Cargo Cult Science'

Fortified by a review of fundamental assumptions, we set about to test a hypothesis from political theory empirically, using data mining methods on historical textual data. We take as our hypothesis the general belief, widely held in the humanities, that political beliefs and the use of figurative language are importantly if not fundamentally correlated.

4.1 Lakoff's hypothesis: politics and metaphor

Metaphor is deep in our language, deep in our concept schemes, or so claim thinkers as different as Jacques Derrida and George Lakoff. Richard Rorty puts it with point: 'It is pictures rather than propositions, metaphors rather than statements, which determine most of our philosophical convictions' (p. 12). Indeed, in two recent bestselling books on moral politics and issue framing, the linguist George Lakoff has claimed that political debates are contests between root conceptual metaphors (Lakoff, 2002). Party affiliation is rooted in metaphorically structured mental models (in 'pictures' not 'propositions'). Lakoff's is a provocative idea, one that has been recently entertained by political strategists influential in the Democratic Party of the United States, and in the past few years, Lakoff has been invited by Nancy Pelosi to coach the Democratic caucus and has consulted with Howard Dean, John Kerry, and Hillary Clinton. However, the validity of the conceptual link between metaphorical structures and political affiliation has not yet been

experimentally verified. Many commentators remain dubious about Lakoff's claims.

Lakoff, a cognitive linguist, bases his claims in the theory of 'conceptual metaphor'. Lakoff believes that our concept schemes are metaphorically structured and argues that metaphors 'play an enormous role in characterizing our worldviews' (p. 63). Linguistic metaphor is evidence of mappings or correspondences between concepts across conceptual domains. 'It is extremely common for such metaphors to be fixed in our conceptual systems, and thousands of such metaphors contribute to our everyday modes of thought' (p. 63). In political contexts, Lakoff argues that moral thinking 'depends fundamentally on metaphorical understanding' (p. 41).

4.2 Hypothesis testing

Lakoff's claims indicate a ground for historical inquiry. If metaphors signal structures of belief and different parties form around different metaphorical models, we should be able to distinguish between different eighteenth-century political parties by distinguishing the kinds of metaphors they used. Thus, we would expect, if Lakoff is correct, that metaphorical usage alone would be sufficient information to predict an author's political affiliation. If Lakoff's claims for metaphor are unfounded, metaphors alone will prove insufficient for this prediction.

We can put this hypothesis to the test by using supervised machine learning methodology. The basic idea is to train a classifier on a set of labeled data, where metaphors are examples and political affiliation are labels, and then to test this classifier on a previously unseen set of metaphors. If the classifier is able to predict the political affiliation of authors based solely on their metaphorical usage, we can take that as supporting evidence for Lakoff's hypothesis.

4.2.1 *Methodology*

We select SVMs as our classifier of choice for these experiments, as this is considered the state of the art classification method for textual data (Joachims, 1998). We give a brief overview of this methodology here; the reader is directed to Scholkopf and Smola (2002) for a complete discussion of SVMs and their variants.

In the standard machine learning approach to text classification, each piece of text is represented as a point in a particular kind of space, which we will describe below. The pieces of text used for training are each given a *label*, generally +1 if it belongs to one of the classes, and −1 if it belongs to the other class.[6] A *linear classifier*, such as a linear SVM, finds a line (or, more properly, a *hyperplane*) that attempts to divide texts of one class from the other.[7] When the two classes are *well separated*, drawing such a dividing line is easy and the classifier is likely to have high predictive accuracy on other texts. When the two classes are not easily distinguished in the space—perhaps because of overlapping or intermingling—then it becomes impossible to draw a 'perfect' dividing line. In this case, the learning method tries to minimize error as much as possible.

As mentioned above, each piece of text is represented as a point (or *vector*) in a certain kind of space. Recall that a space is defined as a set of axes, or coordinates. The text space has a unique coordinate for each of the words in the lexicon, and is thus a *high-dimensional* space. Interestingly, in high-dimensional spaces, it is surprisingly easy to find ways to divide one class from another by taking advantage of many different coordinates. Intuitively, however, such a line has inherently high complexity, and thus may be an over-specific fit to the training data with poor ability to generalize. This notion harkens back to Occam's razor from the fourteenth century, and is supported by statistical learning theory (Scholkopf and Smola, 2002).

SVMs handle the problem of high dimensionality by preferring a solution that both minimizes training error *and* is as simple as possible. This is what is meant by the statement that SVMs *maximize the margin* between the two data classes (Scholkopf and Smola, 2002). Finding a line with the largest possible 'border' between the data classes does both tasks of minimizing error and maintaining simplicity. Typically, the goals of simplicity and accuracy are in opposition; thus, SVMs have a user-set parameter C that determines how much weight to assign to each of these goals. Thus, SVMs are a strong

choice for text classification, which is an inherently high-dimensional task.

4.2.2 *Data*

Our source of metaphorical data was the Mind is a Metaphor database, a hand-curated collection of thousands of metaphors of the mind that occur in eighteenth century English literature (Pasanek, 2006). To prepare our data, we mapped each piece of metaphorical text to an example vector using the binary bag of words approach (Salton and Buckley, 1988). Each word corresponds to a unique coordinate (or, *feature*), and that feature is given a score of 1 if the word occurs in the piece of metaphorical text and 0 otherwise. This approach was chosen over a count-based scheme because the pieces of text were quite short; typically one sentence or a small group of sentences. (We follow Lakoff and a majority of Anglo-American philosophers of language in treating metaphor as a sentence-level phenomenon.)

Each example was given a label of political affiliation of Whig, Tory, or Radical. This labeling was accomplished by labeling the politics of the authors as specified by the *Oxford Dictionary of National Biography* (Harrison, 2007). In those cases where the ODNB did not indicate an author's politics, we threw out the metaphors associated with the author. Furthermore, we availed ourselves of an oversimplification: for the purposes of our experiments, all metaphors produced by Tory author were labeled as 'Tory' metaphors. In all our experiments, we used the *libsvm* package as our SVM engine, with a linear hypothesis and the cost parameter C set to the default setting.

4.2.3 *Cross-validation by metaphor*

In our first trial, we created training data by mapping metaphors to bag of words vectors, using binary feature scoring. No word stemming was performed, and no stop words were removed. We performed 5-fold cross-validation by metaphor—that is, each metaphor was assigned uniformly at random to one of five distinct test sets, with the remainder of the data used as the corresponding training set for that run. We performed random sub-sampling of the data weighted by class size, to ensure that each of the three classes were of equal sizes in our final data set. We report the results for this trial in Table 1, which gives a confusion matrix of predictions compiled over this set of cross-validation tests.

Table 1 Confusion matrix with balanced class sizes, cross-validation by metaphor, no word stemming, stop word removal, or feature selection

	Predicted Whig	Predicted Tory	Predicted Radical	Recall
Actual Whig	434	134	68	0.69
Actual Tory	149	456	56	0.66
Actual Radical	76	83	486	0.72
Precision	0.62	0.70	0.69	Inf. Gain: 0.69

The confusion matrix reports several pieces of information. First, within the grid, the raw counts show how many times an actual Tory, say, was (wrongly) classified as a Whig. Second, the *precision* measure answers the question: of all the times we predicted an example to be a Tory (for example), how often were we correct? Third, the *recall* measure addresses the complementary question: of all the Tories in the data, what fraction did we correctly detect? Finally, *information gain* gives a measure of how much better than random guessing the classifier's predictions are, where pure random guessing scores a 0 (Duda *et al.*, 2000).

Looking at these results, we see a strong diagonal, good precision and recall (against an expected baseline of 0.33), and strong information gain. Thus, based on the results of this experiment, we can say there is a strong signal connecting metaphorical usage to political affiliation, and while correlations should not be confused with a structure of causation (is it, in fact, our politics that predispose us to particular metaphors or does another, more fundamental factor correlate party and metaphor?), it would seem that an author's choice of metaphors signal his or her party affiliation.

In short, Lakoff appears to be right.

4.2.4 *Cross-validation by author*

If we only sought the barest proof for Lakoff's hypothesis, we would be done. However, it is possible that cross-validation by metaphor may not be the most meaningful test for this data. Recall that

the process of cross-validation breaks the data into partitions of test data and training data. With cross-validation by metaphor, standard practice in machine learning is upheld as the same example never appears in the same test and training sets. However, it is possible for some metaphors from a given author to appear in a training set, and other metaphors from that same author to appear in the test set. In this case the classifier may be simply learning to recognize individual author vocabularies rather than an overall political affiliation.

To guard against this, we repeated the experiment with the same experimental design, but this time performing cross-validation by author. That is, all of a given author's metaphors were assigned to the same randomly selected test set.

The results for this experiment are given in Table 2 and show, strikingly, that the signal connecting metaphorical usage to political affiliation has effectively disappeared. Precision and recall are now near the random baseline of 0.33, and there is near zero information gain.

In short, Lakoff appears to be wrong. And where Lakoff is wrong, another student of politics may be right. The historian Lewis Namier and his modern inheritors may come forward here. It is Namier and his followers who argue that the great Parliamentary battles of the mid-eighteenth century were not between Whigs and Tories, but between groups of Whigs. Ultimately, Namier's interpretation of eighteenth-century party politics sweeps party distinctions aside. Namier draws his picture of man from psychoanalysis: instinct, self-interest, habits, and the reproduction of Oedipal structures constitute the deep structure of individuals, nation, and history. By Namier's lights, the 'Whig interpretation' of the eighteenth century focuses intently on ideology and party relations. Namier's corrective replaces the party with the individual and his ambitions.

Table 2 Confusion matrix with balanced class sizes, cross-validation by author, no word stemming, stop word removal, or feature selection

	Predicted Whig	Predicted Tory	Predicted Radical	Recall
Actual Whig	249	211	198	0.38
Actual Tory	275	199	167	0.31
Actual Radical	256	139	250	0.39
Precision	0.32	0.36	0.41	Inf. Gain: 0.04

4.2.5 *Cross-validation by author with basic feature preprocessing*

With conflicting results from our first two experiments, we revisited our assumptions about how best to represent the text. The first two experiments were performed on 'raw' bag-of-words features. That is, each unique word was considered to be a unique feature, so that the words 'mind' and 'minds' would be thought of as distinct features. Furthermore, all words were considered to be features, including very frequent *stop words* such as 'the' and 'what', and very rare words that may have occurred only once or twice in the entire corpus. Treating word-based features in this way imposes little experimenter bias, but may also subject the classifier to unnecessary noise in the data.

It is common practice in text mining and information retrieval to perform word stemming and remove stop words and very uncommon words for improved performance. We follow this practice in this, our third experiment. But the reader must be reminded that, in this experiment we are now removing a great number of little three- and four-letter words—words like 'mine' that our Dickinson scholar found so very provocative. These words that are not readily visible to the literary critic are the centerpiece of studies that follow John Burrows' classic work with Jane Austen's stylistics (Burrows, 1987).

The results for this experiment are given in Table 3, and show some measurable (but not particularly strong) signal connecting metaphorical usage to political affiliation.

Indeed, it may be that neither Namier nor Lakoff is right. These results begin to suggest a new interpretation: the classifier shows better precision in identifying Tories and radicals. Indeed, the results go some way to illustrating J. G. A. Pocock's claim that the outsider politics of Tories and radicals and Opposition Whigs draw on a common language of 'civic humanism' (Pocock, 1985).

Table 3 Confusion matrix with balanced class sizes, cross-validation by author, with stop words and infrequent words removed, and with word stemming

	Predicted Whig	Predicted Tory	Predicted Radical	Recall
Actual Whig	327	179	148	0.50
Actual Tory	266	300	81	0.46
Actual Radical	323	121	201	0.31
Precision	0.36	0.50	0.47	Inf. Gain: 0.16

4.2.6 *Cross-validation by author, with preprocessing and latent semantic analysis*

The word stemming from the third experiment allowed us to capture the similarity between morphological variations of the same word. However, it did not let us take *semantic* similarity into account, such as may exist between, say, 'stone' and 'rock'. Thus, there may have been similar metaphorical usage that was missed in the our previous experiments because of these semantic considerations.

Latent semantic analysis (LSA) is a well-established technique from information retrieval (Skillicorn, 2007) to automatically detect semantic relationships between words, based on detection of co-occurrence via singular value decomposition of the term–document matrix. We perform LSA using the top 50 singular values from our term–metaphor matrix, and repeat the experiment with this semantically informed feature representation.

The results for this experiment are given in Table 4, and now show an increased signal significantly stronger than random, but by no means absolute or definitive.

As we have said all along, Lakoff is more or less right.

4.2.7 *Distant readings and close readings*

When working with metaphors it is important to prepare the data in a way that shows off semantic similarities and ignores stylistic issues. There is a case to be made for the word stemming and LSA: metaphor would seem to be a matter of semantics. And so we massage our data. But clearly, we cannot draw meaningful conclusions from any one of these four experiments in isolation. In order to interpret these results, a *close consideration* of the full set of experimental results is required.

Table 4 Confusion matrix with balanced class sizes, cross-validation by author, with stop words and infrequent words removed, with word stemming, using LSA with top 50 singular values

	Predicted Whig	Predicted Tory	Predicted Radical	Recall
Actual Whig	318	199	157	0.54
Actual Tory	226	337	88	0.52
Actual Radical	268	74	303	0.47
Precision	0.39	0.64	0.55	Inf. Gain: 0.31

One of the ironies here is that machine learning methods, which seemed so promising as a way of performing what Moretti calls *distant reading* or what Martin Mueller calls, perhaps even more provocatively, *not-reading*, is that they require us to trade in a close reading of the original text for something that looks like a close reading of experimental results—a reading that must navigate ambiguity and contradiction. Where we had hoped to explain or understand those larger structures within which an individual text has meaning in the first place, we find ourselves acting once again as interpreters. The confusion matrix, authored in part by the classifier, is a new text, albeit a strange sort of text, one that sends us back to those texts it purports to be about. Like Scott's 'typical' or 'average' hero, Waverley, the table of results stands for a greater sociopolitical history.

4.3 Clustering

One might think that the ambiguities of the previous set of experiments were owing to the employment of a supervised learning methodology that was employed. These supervised classifiers were trained on data using particular choices of data representation and experimental design, choices that encoded implicit assumptions on the part of the experimenters. Perhaps an *unsupervised* learning methodology, such as unsupervised clustering, would allow the data to be expressed with less imposition of practitioner bias or 'fore-projection'.

To explore this idea, we now examine Lakoff's proposed connection between metaphorical usage and political affiliation involved hierarchical clustering of a subset of metaphors. Hierarchical clustering is one unsupervised clustering method which

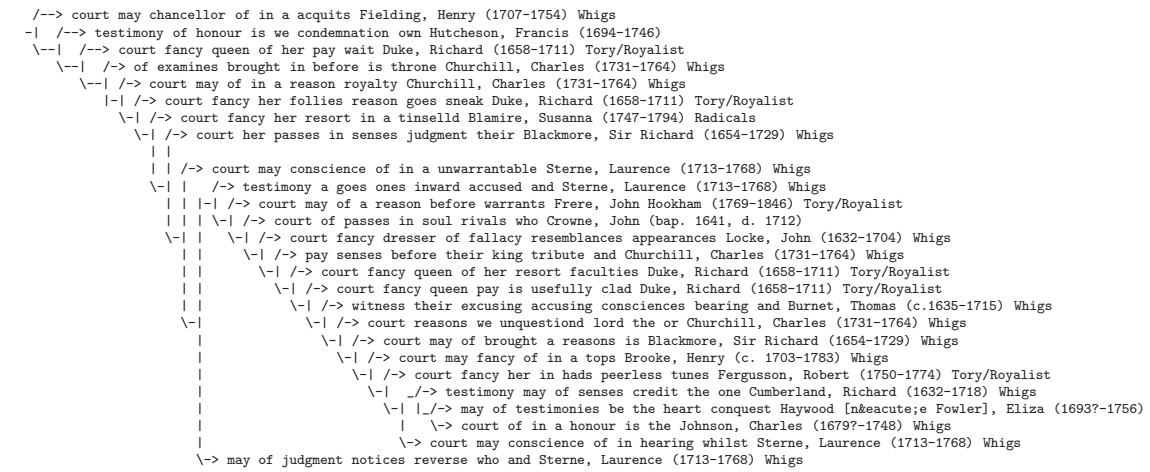

Fig. 1 Clustering court metaphors of the mind with binary feature scoring

attempts to find underlying structure in data, and present it in an interpretable visual format as a hierarchical tree. In general, hierarchical clustering works by iteratively joining the most similar pair of examples; however, there are important design choices to be made in choosing a particular hierarchical clustering algorithm. What is the feature space that will be used, and what similarity measure will be used to determine what the 'most similar' pair of examples is at any given step? As stated above, owing to the No Free Lunch Theorem, it is impossible to choose the best feature representation, scoring method, or similarity measures a priori in the general case. Thus, responsibility falls back upon the human researcher to decide among the available design parameters.

To illustrate this difficulty, we constructed three hierarchical clusterings of all of the 'court' metaphors in the Mind is a Metaphor database (Pasanek, 2006), by representing them with bag of words vectors using three different feature-scoring methods. These feature-scoring methods are:

- Binary scoring—each bag of words vector element was given a score of 1 if the corresponding word occurred in the metaphor, and 0 if it did not. This scoring method considers all words to be equally informative.
- TF–IDF scoring—each word is given an IDF score, which weights rare words as being more informative (Salton and Buckley, 1988), where 'rarity' is measured with respect to a reference corpus (here, the entire metaphor database). This score is then multiplied by a TF count of how many times the word occurred in the metaphor. Thus, rare words occuring often in a metaphor receive high scores.
- Kullback–Liebler divergence (KLD) scoring—here, each word is given a score based on the divergence in probability distribution (Duda *et al.*, 2000) between the word in normal usage (defined by a reference corpus), and the probability distribution of the word in metaphorical usage (defined by the particular subset of metaphors under consideration). Thus, words that are especially strongly connected to the metaphors at hand get high scores.

We then performed hierarchical clustering using the software developed by Stolcke (1996) on the data using each of these three feature representations.

The results for binary feature scoring are shown in Fig. 1 and show no meaningful political grouping of authors based on metaphorical data. The results for feature scoring using the TF–IDF scoring method are given in Fig. 2, and show some political grouping of authors by metaphorical usage. Finally, the results in Fig. 3 show the effect of feature scoring using KLD, which gives very clear groupings

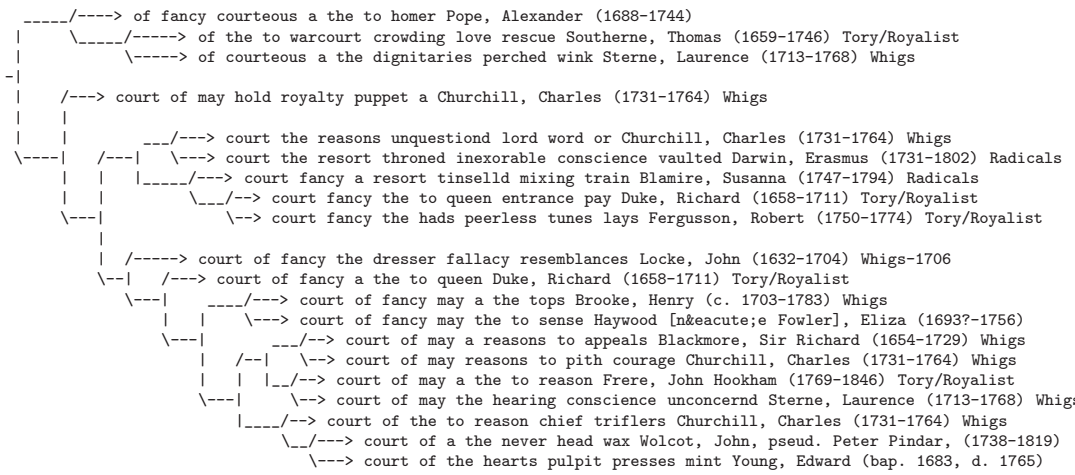

Fig. 2 Clustering court metaphors of the mind with TF–IDF feature scoring

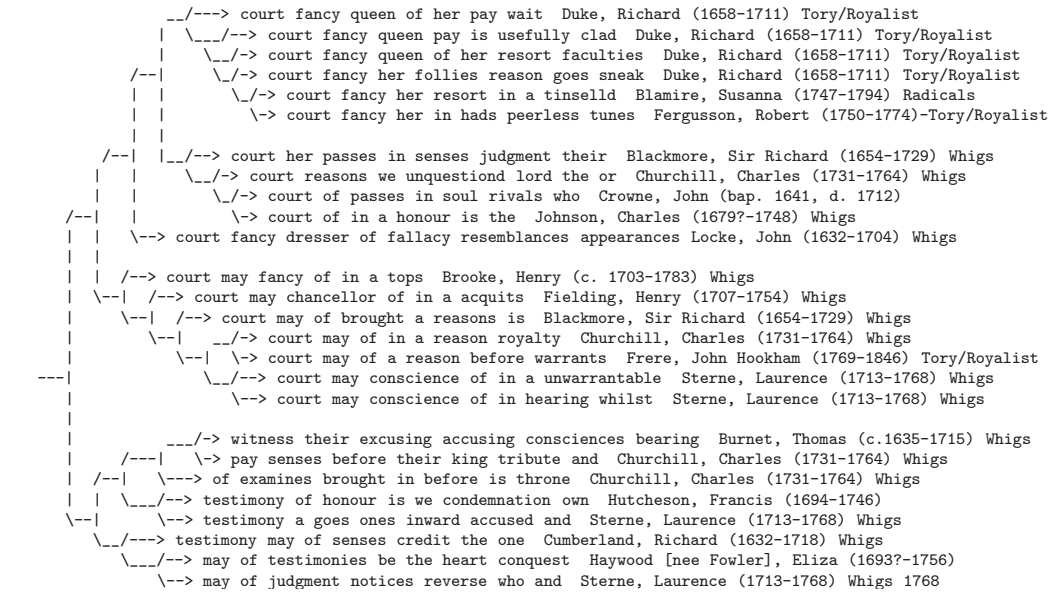

Fig. 3 Clustering court metaphors of the mind with Kullback–Liebler distance feature scoring

of authors into political divisions based on metaphorical usage, and also clearly disambiguates two senses of the word 'court'. Whigs and Tories are clustered away from each other by this word sense disambiguation and we begin to see that a Whiggish political idiom has more to do with testimony and jurisprudence whereas a Tory idiom depends on the Queen's court—Queen Anne's court almost certainly, for it is Queen Anne who reigns during those brief years in the early eighteenth century when a successful Tory ministry controlled the Parliament.

From a data mining perspective, each of these experiments was performed in a valid manner.

Yet each of them conveys different meaning to the literary historian who is trying to establish clear correlations in the eighteenth century between metaphorical usage and political affiliation. Is there a Whig idiom? A constellation of Tory metaphors? Is a political ideology a tissue of commonplaces? These are hotly debated questions in literary criticism and political history. Unfortunately, our experiments do not position us to make any final pronouncement concerning Lakoff's theories. Our experiments do invite new forms of reading and interpretation, as the literary historian is forced to perform a careful study of the results, in each case bringing his own understanding of the period's political rhetoric to bear on these conflicting sets of valid results.

5 Methods and Meaning: a Dada Parable

> *Carefully cut out each of the words that make up this article and put them all in a bag.*
> *Shake gently.*
> *Next, take each cutting out one after the other.*
> *Copy conscientiously in the order in which they left the bag.*
> *The poem will resemble you.*
>
> – Tristan Tzara, 'How to Make a Dadaist Poem'

One difficulty that we encounter repeatedly in data mining for the humanities is that just because results are statistically valid and humanly interpretable does not guarantee that they are meaningful. Consider, as an emblem of our methods, the Dadaist approach of cutting a text into words and then pulling the words at random from a bag.[8] This approach to poetry caused a riot in the 1920s, and yet is based in a statistically valid (if simplistic) methodology known as Gibbs sampling (Duda *et al.*, 2000).

To illustrate this point, we performed Gibbs sampling on the words in the Mind is a Metaphor database to create ten summary lines of fifteen words each, shown in Fig. 4. Taken as short summaries of the database, they are indeed interpretable, especially if we imagine that Gibbs is a poet—an eighteenth-century scribbler of nonsense with a cultivated interest in metaphors for the mind. His poetic lines are a version of those we find in the cluster trees. While they may appear meaningless, we find, with practice, that we can give a gloss or a paraphrase for all varieties of nonsense.

The sixth line—'stamp the resembling and regiment making it reason and it had that goodness inlet unfruitful'—has a decidedly epistemological flavor and is available for interpretation. Let us improvise some punctuation and syntax and borrow a stop word or two as necessary. First, we note that

- fancy irish but is adoration breast now love shake my educated not mind recollect treasures
- soul a the death its the which remember object is of body his its under
- in to scholar the seat lest grown must had sung shall man mind sciences inclination
- the about the steel skull smooth of hang to harden of hell all of the
- for the being of part breast passions in e william of they only render if
- stamp the resembling and regiment making it reason and its had that goodness inlet unfruitful
- hearts an tis of them knowledge sway impartial strongly for paves of eye by mirror
- not snared to and cinders his first very resign magnetised beams balance can former and
- make to and of heels be of wonders with virtue render just persuade self to
- say it stamp one while in the shall to where could rude affections the fear

Fig. 4 Gibbs sampling from eighteenth century metaphors of the mind

stamping metaphors are used by eighteenth-century philosophers to link entities to sense impressions. ('Impression' is a notable metaphor for the mind that connotes anything from wax tablets to printing presses.) In the *Essay on Human Understanding*, Locke distinguishes between the primary qualities of 'Solidity, Extension, Figure, and Mobility', which are 'utterly inseparable from the Body' (II.viii.9) and the secondary qualities ('Colours, Sounds, Tasts, etc.') 'which in truth are nothing in the Objects themselves, but Powers to produce various Sensations in us by their primary Qualities' (II.viii.10).[9] From these disorderly 'unfruitful' resembling impressions which pour through an inlet into the mind, we draft a mental 'regiment'. The stamped impressions of sense are recruited as material for the reason or the understanding to operate upon (the line may now have a Kantian flavor).

The third line, 'in to scholar the seat lest grown must had sung shall man mind sciences inclination' could almost serve as an epigraph to this essay. We leave it to our readers to construct a complete interpretation of the line on their own.

Reading these lines makes for a parlor game or a display of interpretive virtuosity. In truth, what is not interpretable by the literary critic? Nature is a divine book, legible to all who will read it, and, likewise, any text may be naturalized by the efforts of interpretation. The interpreter approaches the artifact and discovers it to be patterned, pregnant with purposiveness, meaningful in advance of his or her expectations. Yet, it would be difficult to present these new texts to the community of literary scholars as hard evidence for scholarly research. Thus, this extreme case serves as a cautionary example. Interpretability alone cannot be our guide in evaluating results from data mining methods applied in humanities. The literary critic—flying the banner 'mine'—can interpret anything, but such interpretation may well be another name for overfitting the data.

6 Conclusion: Standards of Automated Evidence

Literary critics and machine learning practitioners both have well-defined standards that must be met before a new finding is considered to be a trustworthy result. But when we cross discipline, these standards must be revisited. In collaboration, we discover an opportunity to examine our practices. Willard McCarty writes, 'Computational form, which accepts only that which can be told explicitly and precisely' proves 'useful for isolating . . . tacit and inchoate' knowledge (p. 256). Collaborators are forced to set out a program in detail, one that is mutually comprehensible but also one that delivers results that are simultaneously meaningful in two disciplines.

6.1 Technology as proof

The temptation in applying machine learning methods to humanities data is to interpret a computed result as some form of proof or determinate answer. In this case, the validity of the evidence lies inherent in the technology. This can be problematic when the methods are treated as a black box, a critic *ex machina*.

Indeed, we will not devise better methods or better models that sever theory from observation. Circularity is, if we believe Gadamer, at the heart of interpretation. The act of interpretation begins in 'fore-projection' and we bring our linguistic competence, theories, sense of history to bear on whatever text it is we choose to read. The critic who works with a computer scientist will not escape the hermeneutic circle by employing an automated technique to identify patterns that somehow lie await in the texts.

But perhaps this circularity may be framed as a strength and not a weakness of computational methods. Steve Ramsay, for one, explicitly celebrates circularity and warns the literary critic that there are no scientific solutions to interpretive problems. Ramsay's is an admirable embrace of the ineluctable.[10]

Ramsay writes, software 'cannot be neutral . . . since there is no level at which assumption disappears. It must rather, assert its utter lack of neutrality with candor, so that the demonstrably non-neutral act of interpretation can occur' (p. 182). He describes his graphs not as 'objective data' useful for adjudicating some 'humanistic problem'—they are not 'concrete evidence to

support or refute hypotheses or interpretations' (Ramsay, 2005, p. 183).

Machine learning delivers new texts—trees, graphs, and scatter-grams—that are not any easier to make sense of than the original texts used to make them. The critic who is not concerned to establish the deep structure of a genre or validate a correlation between metaphor and ideology, will delight in the proliferation of unstable, ambiguous texts. The deferral of meaning from one computer-generated instance to the next is fully Derridean.

6.2 Hypothesis testing

But what of the critics who are interested in privileging explanation over interpretation? When data mining methods are used for hypothesis testing, we must deal with questions of proof explicitly. How much signal is needed before we consider the experiment to agree with a hypothesis? Clearly, demanding 100% accuracy in the prediction of political affiliation based on metaphorical usage seems to be too strict a test—but where do we draw the line? And what are we to make of partial results, such as the ones in our case study, which show a statistically significant signal that is far from 100% reliable?

Obviously, when the results of an experiment agree with the hypothesis, we cannot consider this to be a *proof* of the hypothesis, only a piece of evidence in its support. Furthermore, we must take care to consider if the connections between data and labels may be coincidental—as appeared to be the case in our cross-validation by metaphor, where the connection found by the classifiers seemed to be more strongly influenced by the connections between individual author identity and metaphorical usage.

Likewise, when the results of an experiment disagree with the hypothesis, we cannot consider this to *invalidate* the hypothesis—these results are only a piece of evidence in the argument against it. And we must consider the possible objections: were the connections not found because the wrong feature space was used, or because the wrong learning method was used, or because insufficient data was employed? Any of these objections could be reasonably used to explain away the importance of a negative result.

The opportunity for self-reflection is welcome. We read Lakoff against the political history of the eighteenth century. Competing results offer opportunities to make decisions between competing explanations.

6.3 Recommendations for best practice

We believe the standards of evidence for the use of data mining in the humanities to be, if anything, more difficult to meet than in data mining for the sciences where there is a clear objective function. Our task is one of quantifying ambiguities and subtleties. To this end, we hope that the community of those involved in data mining for the humanities will converge on a consensus for best practices in this work. We contribute to this collective conversation with the following set of recommendations. Needless to say, these are in addition to the standards of best practice in data mining (such as the use of separate training and testing data, etc.) that are laid out carefully by Salzberg (1997).

6.3.1 *Make assumptions explicit*

Literary data mining is most often a collaborative effort between individuals coming from different backgrounds, who, perhaps, share little common vocabulary. It is incumbent on both literary scholars and data mining experts to make all of their underlying assumptions about the text and its representations explicit. It is these most basic conversations about the nature of a text that we have found most edifying in *our* collaborations.

6.3.2 *Use multiple representations, methodologies*

As we have seen in the case study reported here, different representations of the data highlight different aspects of a text and address different issues in interpretation. Because there can be no single 'best' representation or experimental design, it is important for researchers to consider multiple viewpoints and not be content with a single early result.

6.3.3 *Report all trials*

In data mining for the humanities, results may differ significantly for alternative representations or methodologies. It is important to report all results to present the complexities and underlying issues of the data at hand, rather than cherry picking one or two results from a set. Indeed, in humanities data mining, 'failed' experiments can often be even more informative than 'successful' ones. We refer the reader again to Ramsay's interest in anomalous and outlier results.[11]

6.3.4 *Make data available and methods reproducible*

Experimental results should be verifiable by independent researchers. In general, this means data should either be made available to other researchers upon request, or preferably be posted in an online archive. Even when copyright restrictions make full data disclosure impossible, it may be possible to distribute obscured forms of the data, such as feature vectors rather than full text.

6.3.5 *Engage in peer review of methodology*

As a community, we must foster critical discussion not only of text and literary interpretation, but also of experimental method and interpretation of results. In many ways, this is a more difficult burden of proof than is required in scientific data mining, where objective performance measures allow clear comparisons of approaches. As a community, we need to find an articulate consensus on meaningful standards for experimental evidence provided by data mining.

Ultimately, the goal of data mining in the humanities is not to turn the study of literature into some sort of pseudo-science, but rather to enable scholars to highlight the varying facets of complex issues. We reject overly simplistic interpretations of any form, whether drawn directly from a text or from experimental results. We believe the work of data mining in the humanities is most often about highlighting ambiguities and conflicts that lie latent within the text itself, and it is to be expected that such work will often stall out in inconclusiveness. Indeed, the virtue of automated analysis is not the ready delivery of objective truth, but instead the more profound virtue of bringing us up short, of disturbing us in our preconceptions and our basic assumptions so that we can exist, if only for a moment, in uncertainties, mysteries, and doubts. Should we learn to forestall interpretation we may come to revise our prejudices, theories, and fore-projections in terms of what emerges.

References

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Oxford University Press.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification (2nd Edition)*. New York: Wiley-Interscience.

Gadamer, H.-G. (2000). *Truth and Method*. (Weinsheimer, J. and Marshall, D., Trans.) 2nd Rev. edn. New York: The Continuum Publishing Company, p. 175.

Harrison, B. (ed.) (2007). *Oxford Dictionary of National Biography, Online Edition*. Oxford University Press. www.oxforddnb.com.

Joachims, T. (1998). *Text Categorization with Suport Vector Machines: Learning with Many Relevant Features*. *ECML '98: Proceedings of the 10th European Conference on Machine Learning*.

Lakoff, G. (2002). *Moral Politics: How Liberals and Conservatives Think*. Chicago: University of Chicago Press.

Lancashire, I. (1993). Computer-assisted critical analysis: a case study of margaret atwood s handmaid s tale. In *The Digital Word: Text-Based Computing in the Humanities*.

Lukács, G. (1983). *The Historical Novel*, (Mitchell, H. and Mitchell, S., trans.) Nebraska: University of Nebraska Press.

Metsis, V., Androutsopoulos, I., and Paliouras, G. (2006). *Spam Filtering with Naive Bayes – Which Naive Bayes? Third Conference on Email and Anti-Spam (CEAS)*.

Miall, D. S. (1993). Beyond the word: reading and the computer. In *The Digital Word: Text-Based Computing in the Humanities*.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Pasanek, B. (2006). Eighteenth century metaphors of the mind, a dictionary. Doctoral dissertation. Stanford University.

Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M., and Smith, M. N. (2006). Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces. *International Conference on Digital Libraries. Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*. Chapel Hill, NC. New York: ACM Press, pp. 141–150.

Pocock, J. G. A. (1985). *Virtue, Commerce, and History: Essays on Political Thought and History, Chiefly in the Eighteenth Century*. Cambridge: Cambridge University Press.

Ramsay, S. (2005). In praise of pattern. *Text Technology*, (2).

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**(5): 513–523.

Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, **1**(3).

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

Skillicorn, D. (2007). *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. Boca Raton: Chapman and Hall.

Stolcke, A. (1996). Cluster 2.9. http://www.icsi.berkeley.edu/ftp/global/pub/ai/stolcke/software/cluster-2.9.tar.Z.

Unsworth, J. (1997). Creating digital resources: The work of many hands. *Digital Resources for the Humanities*. http://www3.isrl.uiuc.edu/~unsworth/drh97.html

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, **27**(11).

Wolpert, D. H. and Macready, W. G. (1995). *No Free Lunch Theorems for Search*. Technical report. Santa Fe Institute, Santa Fe, NM, USA.

Notes

1 In an essay on the state of the art, published in the 1993 collection *The Digital Word*, David S. Miall discusses a representative moment of foundational anxiety. The year was 1978. Klaus Schmidt attacked the KWIC concordance itself, that early darling of computer-assisted textual analysis. A concordance is not useful, complained Schmidt, to the scholar who hopes to make a discovery, because the concordance demands of the scholar that he already know what he is looking for. Its efficacy as a textual instrument lies in its locating moments in the text the user of the concordance knows he wants to find.

2 'Mine' is a key feature in Dickinson's erotics, or so the classifier teaches us. One wonders who is in on the joke here: the Dickinson scholar? Kirschenbaum? the reader? The classifier? After all, the machine identifies 'mine' as an erotic word!

3 See Matthew Kirschenbaum's 'Poetry Patterns, and Provocation: The NORA Project. *The Valve: A Literary Organ*, 12 January 2006. See also the more formal report authored by Kirschenbaum and others (Plaisant *et al.*, 2006).

4 Gadamer locates the part-whole dialectic in the earliest hermeneutic writings: 'the whole of Scripture guides the understanding of individual passages: and again this whole can be reached only through the cumulative understanding of individual passages. This circular relationship between the whole and the parts is not new. It was already known to classical rhetoric, which compares perfect speech with the organic body, with the relationship between head and limbs' (Gadamer, 2000).

5 Moretti also lists *Black Forest Village Stories* and *I Malavoglia* as examples.

6 This sort of unsubtle *ground truth* labeling is one area that may make the literary scholar uneasy. Although in this article we use this framework for simplicity, it is worth noting that there are machine learning methods capable of dealing with labels that are not strictly black-and-white, such as probabilistic labelings.

7 When there are more than two classes of data, a *multi-class* SVM is used, which essentially consists of several pairwise classifiers (Scholkopf and Smola, 2002).

8 One thinks also of William Burroughs' cut-up method or Gilles Deleuze's 'pick up' procedure.

9 Primary qualities are like that which produces them; secondary qualities are not (II.viii.15).

10 In his essay 'In Praise of Pattern', Ramsay finds an opportunity to go on a 'fishing expedition' in which after discovering a new and interesting technology he struggled to find a way to apply it (Ramsay, 2005).

11 And we look forward to forthcoming papers co-authored by Shlomo Argamon, Mark Olsen, Russel Horton, Charles Cooney, and Sterling Stein that, likewise, celebrate moments of *mis*-classification.