Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities

D. Sculley *(dsculley@cs.tufts.edu)*
Department of Computer Science
Tufts University
Brad Pasanek *(bpasanek@annenberg.edu)*
Annenberg Center for Communication
University of Southern California

In working across and between disciplines, it is the tacit assumptions that may be most destructive to meaningful collaboration. Ours is a state of mutual ignorance, and the goals and practice of the professional literary historian and the machine-learning researcher are equally obscure. But in collaboration mutual ignorance becomes an opportunity for self-reflection, clarification, and the speaking of what is usually unspoken. Willard McCarty writes, "Computational form, which accepts only that which can be told explicitly and precisely" proves "useful for isolating ... tacit and inchoate" knowledge (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. In this paper, we discuss the tacit assumptions that accompany data set preparation, hypothesis testing, and data exploration in order to deliver prescriptive claims. We propose a communication protocol designed to bring hidden and tacit assumptions into plain view where they may be discussed and analyzed. This paper is the third in a series of collaborative efforts undertaken by the two authors. It is informed by real experience working together: working often at cross purposes, garbling a common language, but ultimately producinresults that are of interest to both computer scientists and literary scholars.

Transforming literary content into data for machine learning methods requires the adoption of a number of initial assumptions, each of which significantly impacts the final results. First, the collaborators must select or design an appropriate data representation. The selection of a bag of words model may be one such decision, but other feature mappings such as parse trees or link structure graphs may be more informative for a given task. Once the textual material is represented, we must decide upon a method of feature weighting; that is, we must decide if some features are more important than others and how much so. Because many learning methods prove intractable when working with very large numbers of features, feature selection is necessary in order to enable computation. Sophisticated, ambiguous, unstable texts must be normalized to make comparisons across texts meaningful—so that the choice of normalization method is critical. So, too, are methods of noise filtering. There are no clear objective choices among methods, because each choice introduces a set of assumptions and biases. We demonstrate these difficulties with experiments on a range of literary data. A loose analogy is here drawn as the literary scholar may choose to cite post-colonial theory to the exclusion of queer theory or practice close reading to the exclusion of historical analysis. We do not argue that structuring assumptions be minimized or eliminated--this is impossible--but we do make the case that in interdisciplinary work especially, it is important for the impact of each assumption to be assessed and reported at the outset. The critic is often a bricoleur, borrowing from literary theory in promiscuous fashion. In preparing data, bricolage is not a ready option and the collaborators must make painful compromises.

The use of machine learning for the testing of literary claims also has several potential pitfalls. The broadest is the impact of the No Free Lunch Theorem, which states that there can be no single machine-learning algorithm that gives optimal performance on all data sets. The choice of a learning algorithm entails, again, the adoption of tacit assumptions about the underlying structure of the data. We may assume that data is linearly separable (which is often a true assumption in text classification), or that the data examples are statistically independent of one another (which is often false in the text domain). As we demonstrate experimentally, these assumptions carry significant impact on the results of the data mining. In the literary domain, selection bias seems particularly problematic as we navigate the politics of canon formation, the difficulty of defining of genre, and the vagaries of influence--all of which trouble the initial selection of texts.

An important question in both machine learning and literature is that of generalization. Do the results and models we discover apply only to our particular data set (as in the case of rote learning), or do these patterns also describe new periods and genres, data we have not yet investigated? In truth, machine-learning methods can never guarantee generalization. However, they do offer statistical bounds on the probability that a model will generalize. According to the Probably Approximately Correct paradigm of computational learning theory, a model that achieves a given level of accuracy on a training data set will likely achieve a predicted level of accuracy on a test data set from the same distribution. The computer scientist emphasizes that generalization bounds are only valid under assumptions of statistical independence in the training data. Care must be taken in the literary domain to ensure that probabilistic assumptions are satisfied. Otherwise, the findings

may reflect little more than the selection bias of the investigator. We provide concrete examples of these issues using data from literary analysis, and give guidelines for determining when a generalization assumption may or may not be valid.

The literary scholar often turns to computational methods to explore large numbers of texts--more texts than one human could ever read closely. In this last case, the scholar may not have a hypothesis to test, but is instead looking for new perspectives on literary history. In a word, the literary scholar hopes to be surprised by the computer scientist. However, surprise is too easy a commodity to supply in data mining. Consider that some of the first literary data miners were the Dadaists and Surrealists, who produced poetry by cutting a printed text nto pieces and pulling those pieces randomly from a bag. In machine learning, this method of textual analysis is known as Gibbs sampling (Duda), and has been used in recent work on probabilistic author-topic modeling (Steyvers). This sort of surprise, however, may not be that which a literary scholar desires--it may not be a meaningful surprise. Thus, the scholar must define for the machine-learning specialist exactly what sort of surprises are desired, so that the appropriate data mining methods may be applied. This is a curious hermeneutic circle--the critic worries that requesting a particular kind of surprise effectively removes true surprise from the process. Data exploration requires a bound on the unknowns to be meaningful and productive. We adduce examples of this need with experiments in anomaly detection on literary data.

Data exploration may be performed by employing data visualization techniques, or by using unsupervised methods of machine learning such as clustering. In both of these situations, it is important to keep the cartographer's dilemma in mind. In order to understand large data sets in high-dimensional space both the literary scholar and the computer scientist require some form of dimensionality reduction. While reductive methods may, indeed, enable new insights, they may also produce artifacts--strange islands analogous to the distorted, massive projection of Greenland on most two-dimensional maps of the world--that give a distorted view of the underlying structure.

Two specific dangers, then, accompany data exploration. The first is that a distorted artifact, a picture, may be mistaken for an underlying truth. The second is that once a data set has been fully explored, it may no longer be valid to use it for hypothesis testing. An exhausted data set prompts us to move on to a new set of texts, to generalize as discussed above. But moving to a new set of data, we often discover that our hypothesis is not portable and fails to generalize. The history of literature is a "collective system," as described by Franco Moretti in Graphs, Maps, and Trees (4), but law-like generalizations are difficult to frame and even harder to transport from text collection to text collection: we discover patterns, yes, and congruent patterns may be discovered in different collections. To say more is to abandon many of the certainties that the literary scholar had hoped machine learning would provide him with. In preparing a data set we construct a system; leaving that data set behind we leave behind its artificial systematicity as well. Here the literary scholar is surprised to find the computer scientist a more thoroughgoing poststructuralist than himself.

It may seem that data mining offers no more claims to objectivity than literary scholarship -- and indeed, from a certain perspective, this is the case. At its worst, data mining and visualization techniques produce mere inkblots that do little more than manifest the hidden (and indeed, perhaps even unknown) biases of the researchers. However, these explorations gain in consequence as the tacit assumptions of both the literary scholar and the data miner are clearly stated. To assist in foregrounding assumptions, we propose a protocol for researchers in these disparate fields. This protocol includes ways of talking about patterns in a common language, for defining meaningful data representations, and for selecting appropriate statistical assumptions. Only when careful preparatory work is done can data mining have a claim to meaning in the humanities.

Bibliography

Duda, R. O., P. E. Hart, and D. G. Stork. *Pattern Classification.* 2nd Edition. Wiley-Interscience, 2000.

McCarty, Willard. "Modeling: A Study in the Meaning of Words." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Malden, MA: Blackwell Publishing Ltd, 2004. 255-70.

Moretti, Franco. *Graphs, Maps, Trees.* London: Verso, 2005.

Salzberg, Steven. "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach." *Data Mining and Knowledge Discovery* 1.3 (1997): 317-328.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. Griffiths. "Probabilistic Author-Topic Models for Information Discovery." *Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining.* New York: ACM Press, 2004. 306-315.