Interpretability in Human-Centered Data Science

David Bamman
School of Information
University of California, Berkeley
dbamman@berkeley.edu

Abstract
This paper advocates for a research direction in human-centered data science focused on *interpretability*, which is often in conflict both with predictive accuracy (more complex, non-linear models are often superior predictors to simpler yet interpretable models) and representational complexity (models with more realistic features are often better fits to data than models with fewer, simpler features). What consequences do these tradeoffs have in practice, and to what degree are they necessary compromises? Can we develop methods that are simultaneously interpretable, highly predictive, and representationally complex?

Author Keywords
Interpretability; transparency; data science

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous; I.5.1 [Pattern Recognition]: Models

Introduction
The statistical models that underlie data science are often used for two ends: making predictions and understanding causal (or more simply, correlational) effects [5]. Many domains have no need for prediction at all, but use the machinery of predictive models to understand the influence of observable features on some outcome. Research in quan-

titative literary studies, for example, can develop a model to predict authorship from observed text, but only rarely for the sake of inferring the author of an unknown text [19]; more common is using such a model to identify characteristic style—the features that discriminate one author from another. Much social scientific work in "predicting" political persuasion, personality, gender, age, and other demographic variables from text and other observed behavior function likewise, where actual predictions are less of interest than the characteristic features that are learned to discriminate between classes [10, 4, 2, 12, 22, 23]

Others make use of prediction but are required (through regulatory or other means) to have transparency in the explanation—such as diagnosing medical conditions or assessing credit risk. For cases where predictive accuracy is the primary concern, the information gained from understanding what a model is learning can be instructive in suggesting new features to include.

In all of these models, there is often a tension between the following desiderata:

- Predictive accuracy. On held-out data (not used to train the model) where some true label is known, how accurate are predictions? Even in cases where predictions are not the primary quantity of interest, high predictive accuracy can still be a good measure of the generalizability of the model.
- Interpretability. To what degree can people understand the mechanism of what's learned, either at the scale of an entire model (what features broadly distinguish class A from class B?) or item-level decisions (why was data point x classified A?).
- Representational complexity. Models are often necessary simplifications of the world, and differ in the richness with which data points are described. What is the appropriate level of description for a given task?

The relationship between predictive accuracy and representational complexity has been well explored through the bias-variance tradeoff (more complex models are inherently less biased than simpler models, but come at a cost of greater variability in predictions), but the interaction between these competing ends and *interpretability* has been less much explored. As data science pushes further into outcomes where humans are involved, we advocate for further work in this direction, and outline several research questions below.

Interpretability

What does it mean for a predictive model to be *interpretable*? First, and most critically, interpretability is not a monolithic phenomenon, but rather depends on the use case; different ends entail very different model designs, and it's important to distinguish between them. For lending institutions subject to the Equal Credit Opportunities Act (ECOA), this may simply be an *enumeration* of input features, providing an audit trail guaranteeing that information concerning race, color, religion, national origin, sex, marital status, or age are not present in predictive models assessing credit risk. For those designing predictive models, this may be a small set of *global* class-level distinctions between those assessed to have high credit risk and those with low risk; for individuals denied credit, this may be the specific elements in their own feature representation that were most responsible for their *local* classification decision.

Predictive models in classification differ very strongly in the degree to which they are conducive to being interpretable by people; models also exhibit significant variation within

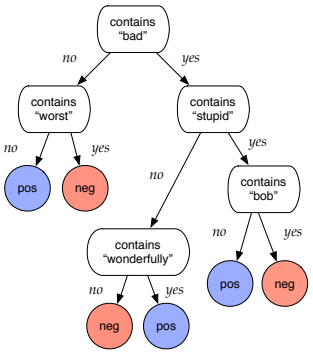

Figure 1: Decision tree for movie sentiment prediction.

feature	coefficient
memorable	0.665
hilarious	0.596
terrific	0.579
others	0.547
excellent	0.546
...	...
waste	-0.724
awful	-0.742
nothing	-0.749
bad	-0.810
worst	-0.815

Table 1: Binary logistic regression coefficients for movie sentiment prediction.

their class as well. To take the simple of example of the binary prediction problem of judging whether a movie review is positive or negative (using data from Pang and Lee [21]), figure 1 illustrates one of the conventionally most interpretable models—a decision tree, which specifies the series of feature conjunctions that result in a classification label. The ability to directly read the sequence of decisions for the model overall and for any instance-level decision makes this model "interpretable"; where this human-level interpretability begins to break down comes with trees of great depth, where classification decisions are attributable to dozens or hundreds of features. Table 1 illustrates the learned model for another traditionally interpretable classifier—binary logistic regression, which learns a weight β_i for each feature x_i and assigns the probability of a class as proportional to $\exp(-\beta^\top x)$. Positive weights in this model correspond to features that are highly indicative of the positive class; negative weights vice versa. With small feature sets, this feature ranking enables interpretability at both a global level (the features at the ends are most characteristic of the two classes learned) and at an instance level (where the features present in a given item being classified can also be ranked). With large feature sets numbering in the tens or hundreds of thousands (as, for example, in classification involving text), interpretability begins to break down; sparse models (such as those involving ℓ_1 regularization) can help make the resulting active feature sets small (and hence more interpretable), but often come at a cost in accuracy.

At the other end of the interpretability spectrum are more complex models that contain non-linear interactions between features. Random forests are aggregations of many small decision trees, each trained on a subset of the data and features; they are far superior predictors to a single decision tree, but come at the cost of interpretability, since a classification decision is now due to hundreds or thousands of local feature conjunctions (one for each decision tree in the forest). Neural networks, especially dense multilayer versions, are often better predictors than simpler linear models, but are notorious for being difficult to interpret, since the impact of any individual feature is spread out throughout the network and often interacts with other features in complex ways.

Aside from the scientific goal of understanding the relationship between features and their dependent effects, or from the insight into new features that model inspection can provide, interpretability is also intimately bound up with *transparency*—while transparency (and its inverse, opacity) takes many forms [6], interpretability is important for problems such as presenting users with rationales for the predictive decisions that impact them. In user design, giving users rationales for algorithmic decisions helps with collaborative filtering [11, 25] and context-aware computing [16, 15], encouraging trust in the algorithmic process by exposing its inner workings. Enabling transparency allows the possibility of giving users control over inferences made about them (letting them "cloak" highly predictive features if they know what those features are) [7].

Q1. How can we formalize "interpretability"?

The first research question we can ask is, at its core, the most important one: how do we formally describe interpretability so that we can measure the degree to which one model is more "interpretable" than another?

To some degree, model selection criteria that penalize the complexity of a statistical model (usually in terms of the number of features, or degrees of freedom, it has)—such as through the ℓ_1 norm or the Akaike/Bayesian information criterion—are a step in this direction, but only apply to a very small subset of available models, and even then only

directly address interpretability as a function of the number of active features, and not, for example, the interpretability of the features themselves, their relationship to the outside world, or the complexity of their interaction within the model. As Freitas [9] points out, model size can be thought of as a "syntactic" description of a model, but does not address a model's semantics—how features combine to create meaning within it (in a way that can be legible to humans).

Others have made progress in identifying features that are influential for predictive outcomes— at the level of individual classification decisions, Martens and Provost [18] define an instance-level explanation to be the minimal set of features that change the classification for an item; others weight features by their differential impact on the class probability, either individually [24] or in sets [26]. At the model level, productive lines of research include both constraining the space of models to encourage interpretability (such as forcing feature coefficients to be integers [28]) or enriching simpler models while preserving their interpretability (such as modeling decision lists in a Bayesian setting [13]).

Each of these methods, however, only addresses one slice of interpretability (feature contribution to decisions) and not the broader sensemaking process by which humans judge models to be interpretable as they use them in decision-making. From a practitioner's perspective, what models are "interpretable" enough to give rise to new knowledge?

Q2. How can we add interpretability to complex models?

A second research question of interest involves balancing the tradeoffs between interpretable (but simpler) models and complex (but uninterpretable) models. Can we develop methods that are simultaneously interpretable, highly predictive, and representationally complex?

Complex models that involve non-linear transformations of input data (such as neural networks with non-linear activation functions and support vector machines with non-linear kernels) tend to be inherently less interpretable than corresponding linear models. Much work has attempted to make these more complex models interpretable by approximating their behavior with simpler models (such as decision trees or rule sets trained not on original training data but rather on the predictions of the more complex model) [8, 1, 17, 3].

Alternative lines of research have attempted to leverage visualization techniques to understand what more complex models are learning; while this is especially pronounced in image recognition [27, 20], recent work has exploited this trend for natural language as well [14].

Both of these approaches necessarily depend on the formalization of "interpretability" outlined in Q1—for the former, in what choice of "interpretable" models are used to approximate the more complex ones; for the latter, the choice of methods to describe their behavior. Formally operationalizing "interpretability" has practical impact here as well.

Conclusion

As data science pushes further and further into the human space, involving people either as the objects of predictive models or the consumers of analytical methods, understanding what predictive models are learning is becoming increasingly important—for establishing audit trails, for suggesting and prioritizing hypotheses to test, and for facilitating the general sensemaking process. There is much work to be done: we need to operationalize "interpretability" in a way that's resonant with our own, human, judgments of the term, and cultivate its use in new models. In doing so, we can create the foundation on which other desiderata—accountability, trust, and transparency—can stand.

REFERENCES

1. Robert Andrews, Joachim Diederich, and Alan B. Tickle. 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems* 8, 6 (1995), 373 – 389. Knowledge-based neural networks.
2. Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12, 9 (2007).
3. Bart Baesens, Rudy Setiono, Christophe Mues, and Jan Vanthienen. 2003. Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Management Science* 49, 3 (2003), 312–329.
4. David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics* 18, 2 (2014).
5. Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16, 3 (08 2001), 199–231.
6. Jenna Burrell. 2016. How the Machine 'Thinks': Understanding the Opacity of Machine Learning Algorithms. In *Big Data and Society*.
7. Daizhuo Chen, Samuel P. Fraiberger, Robert Moakler, and Foster Provost. 2015. Enhancing Transparency and Control when Drawing Data-Driven Inferences about Individuals. In *SSRN*.
8. M.W. Craven and J.W. Shavlik. 1996. Extracting Tree-Structured Representations of Trained Networks. *Advances in Neural Information Processing Systems* 8, 8 (1996).
9. Alex A. Freitas. 2014. Comprehensible Classification Models: A Position Paper. *SIGKDD Explor. Newsl.* 15, 1 (March 2014), 1–10.
10. Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 253–262.
11. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 241–250.
12. Susan C. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10, 4 (2006), 439–459.
13. Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.* 9, 3 (09 2015), 1350–1371.
14. Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and Understanding Neural Models in NLP. In *ArXiv*.
15. Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp '09)*. ACM, New York, NY, USA, 195–204.

16. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128.

17. David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. 2007. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research* 183, 3 (2007), 1466 – 1476.

18. David Martens and Foster Provost. 2014. Explaining Data-driven Document Classifications. *MIS Q.* 38, 1 (March 2014), 73–100.

19. Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *J. Amer. Statist. Assoc.* 58, 302 (1963), 275–309.

20. A. Nguyen, J. Yosinski, and J. Clune. 2015. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Computer Vision and Pattern Recognition*.

21. Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the ACL*.

22. Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 430–438.

23. Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents (SMUC '10)*. ACM, New York, NY, USA, 37–44.

24. Marko Robnik-Šikonja and Igor Kononenko. 2008. Explaining Classifications For Individual Instances. *IEEE Trans. on Knowl. and Data Eng.* 20, 5 (May 2008), 589–600.

25. Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. ACM, New York, NY, USA, 830–831.

26. Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications Using Game Theory. *J. Mach. Learn. Res.* 11 (March 2010), 1–18.

27. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Dumitru Erhan Joan Bruna, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.

28. Berk Ustun and Cynthia Rudin. 2014. Methods and Models for Interpretable Linear Classification. In *ArXiv*.