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.

- 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
their class as well. To take the simple of example of the bi-
ary 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 inter-
pretable 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 interpretabil-
ity 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 lo-
gistic regression, which learns a weight $\beta_i$ for each feature
$x_i$ and assigns the probability of a class as proportional
to $\exp(-\beta^T x)$. 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 fea-
tures 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 in-
volving text), interpretability begins to break down; sparse
models (such as those involving $\ell_1$ 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 be-
tween 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 thou-
sands of local feature conjunctions (one for each decision
tree in the forest). Neural networks, especially dense mul-
tilayer versions, are often better predictors than simpler
linear models, but are notorious for being difficult to inter-
pret, 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 rela-
tionship 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, opac-
ity) 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, giv-
ing users rationales for algorithmic decisions helps with
collaborative filtering [11, 25] and context-aware comput-
ing [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 “interpretable”?
The first research question we can ask is, at its core, the
most important one: how do we formally describe inter-
pretability 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 $\ell_1$ norm or the Akaike/Bayesian informa-
tion criterion—are a step in this direction, but only apply to
a very small subset of available models, and even then only

---

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

<table>
<thead>
<tr>
<th>feature</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>memorable</td>
<td>0.665</td>
</tr>
<tr>
<td>hilarious</td>
<td>0.596</td>
</tr>
<tr>
<td>terrific</td>
<td>0.579</td>
</tr>
<tr>
<td>others</td>
<td>0.547</td>
</tr>
<tr>
<td>excellent</td>
<td>0.546</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>waste</td>
<td>-0.724</td>
</tr>
<tr>
<td>awful</td>
<td>-0.742</td>
</tr>
<tr>
<td>nothing</td>
<td>-0.749</td>
</tr>
<tr>
<td>bad</td>
<td>-0.810</td>
</tr>
<tr>
<td>worst</td>
<td>-0.815</td>
</tr>
</tbody>
</table>
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


