Regulating Inscrutable Systems

March 21, 2017
Black Box
Opaque
Secret
Not Transparent
Unintelligible
Unknowable

Inscrutable
“We cannot effectively regulate what we do not understand”
Barriers to Explanation

• Secrecy
  – Trade secrets
  – Gaming
• Specialized knowledge
• Contingency
• Inscrutability
  – Extreme complexity
  – Semantics
“Explanation” Is Underspecified

• Why did the glass shatter?
  – Was dropped, gravity, glass is brittle, chemical composition, ground is solid, ground is harder, ...

• Context is required and usually inferred
  – Someone upset about cleaning up glass shards vs. chemistry class
  – Without context, explanation mismatch

• Another example: Willie Sutton
How do we rely on explanation in regulation?
Three Layers of Explanations

• **WHAT** happened in the individual decision?
  – Results, inputs, dominant factors, etc.

• **HOW** are the decisions made?
  – Description vs. input-output
  – Full vs. partial

• **WHY** are the decisions made that way?
  – Assumptions, choice of target variable, biases, etc.
  – Must be external
Ex: “Harvard Law” Filter
Connections between the Layers

• If we know **HOW** decisions are made, we know **WHAT** each decision will be.
• If we understand **HOW** decisions are made, we know what questions to ask about **WHY** they were made that way.
• The **HOW** layer is in the driver’s seat.
The Effect of Inscrutability

- Humans can no longer reason about the **HOW** layer
  - Even with full transparency
- Cannot predict **WHAT** layer
- Cannot figure out what we need from **WHY** layer
Existing Law: Credit Scoring and GDPR
1. Credit Scoring
Adverse credit determinations (or other determinations using credit info) require a “statement of specific reasons”

- Prevent discrimination in credit
- Consumer education
- Error checking
Statement of Reasons

• Must be specific

• Must include all principal reasons
  – But “disclosure of more than four reasons is not likely to be helpful to the applicant.”

• Must be the actual reasons
  – E.g., not education as income proxy.
Sample Form Notice (from Reg B)

___Credit application incomplete
___Insufficient number of credit references provided
___Unacceptable type of credit references provided
___Unable to verify credit references
___Temporary or irregular employment
___Unable to verify employment
___Length of employment
___Income insufficient for amount of credit requested
___Excessive obligations in relation to income
___Unable to verify income
___Length of residence
___Temporary residence

___Unable to verify residence
___No credit file
___Limited credit experience
___Poor credit performance with us
___Delinquent past or present credit obligations with others
___Collection action or judgment
___Garnishment or attachment
___Foreclosure or repossession
___Bankruptcy
___Number of recent inquiries on credit bureau report
___Value or type of collateral not sufficient
___Other, specify: ____
Credit Scoring Confounds ECOA

• The statement of reasons works sometimes:
  – Certain reasons, like “unable to verify residence” or “no credit file” are self-explanatory
  – Human credit manager denies for a single reason.
    • Much more common in the 70s
• But credit scores add complexity
The Addition of Complexity

- Scoring bases decision on point total, so many factors all matter at once
- Factors are non-monotonic and appear arbitrary, so difficult to explain
- Thus, it in an inscrutable system.
Credit Scoring – Only the WHAT

- FCRA/ECOA/Reg B only asks for reasons regarding a specific decision
  - No information about HOW the points are assigned
  - No information about WHY the points are assigned that way
2. General Data Protection Regulation (GDPR)
General Data Protection Regulation

• Ongoing debate about “right to explanation”

• Articles 13-15 call for “meaningful information about the logic involved”

• What does *that* mean?
  – No one really knows yet
  – Changed from “knowledge of the logic involved” in Data Protection Directive
Like ECOA, but different

• Specific decision vs. logic of the system
  – WHAT vs. HOW
  – “Meaningful information about the logic” seems to be a call to repair inscrutability of HOW layer
• But in practice, not always clearly separable
• Still doesn’t seek normative explanation
Summing Up

• Credit Scoring asks for **WHAT**
• GDPR asks for **HOW**
• Other sources are required for the **WHY**
• Two problems:
  – Complexity of causation might mean things are not explainable in reality, and making them so reduces accuracy. Therefore human explanation = bias
  – Not clear this is true
Interpretability Overview
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables \( x \) (independent variables) go in one side, and on the other side the response variables \( y \) come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

\[ y \leftarrow \text{nature} \leftarrow x \]

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;

Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from response variables = \( f(\text{predictor variables, random noise, parameters}) \)

\[ y \]

Model validation. Yes–no using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function \( f(x) \) — an algorithm that operates on \( x \) to predict the responses \( y \). Their black box looks like this:

\[ y \leftarrow \text{unknown} \leftarrow x \]

Model validation. Measured by predictive accuracy.

Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;
Four Categories

- Favoring Interpretable Methods
- Global Explanations
- Explaining Specific Decisions
- Task Specific Techniques
Great for Compliance?

• Favoring Interpretable Methods
  – Just works!

• Global Explanations
  – Great for GDPR!

• Explaining Specific Decisions
  – Great for ECOA!

• Task Specific Techniques
  – Not all that useful for general regulation.
Is the Trade-Off a Problem?

• Not all methods necessarily have a straight trade-off
  – E.g., Avoiding overfitting helps interpretability and accuracy
• Must weigh normative goals
  – Accuracy vs. explanation
  – Perhaps a minimum threshold of explanation is simply required
Interpretability and the **WHY**

- Cannot illuminate the **WHY** layer
  - Still need to ask questions of the design
- But *can* connect the design decisions to how it ultimately works
- Not sure what we will ultimately find objectionable, but now we can ask
Limits of Interpretability

- Cannot resolve normative disagreement
  - What if the “why” really is “patterns in the data?”
  - Disagreement about what even counts as discrimination
Let’s Also Explore Other Options

• Directly ask about the **WHY**
  – Credit: creditworthiness or maximum profit?
• Tools that just fix the problem
  – E.g. Discrimination-aware data mining
• Regulations that just fix the problem
  – Draw on some environmental law?