Auditing Black Box Models

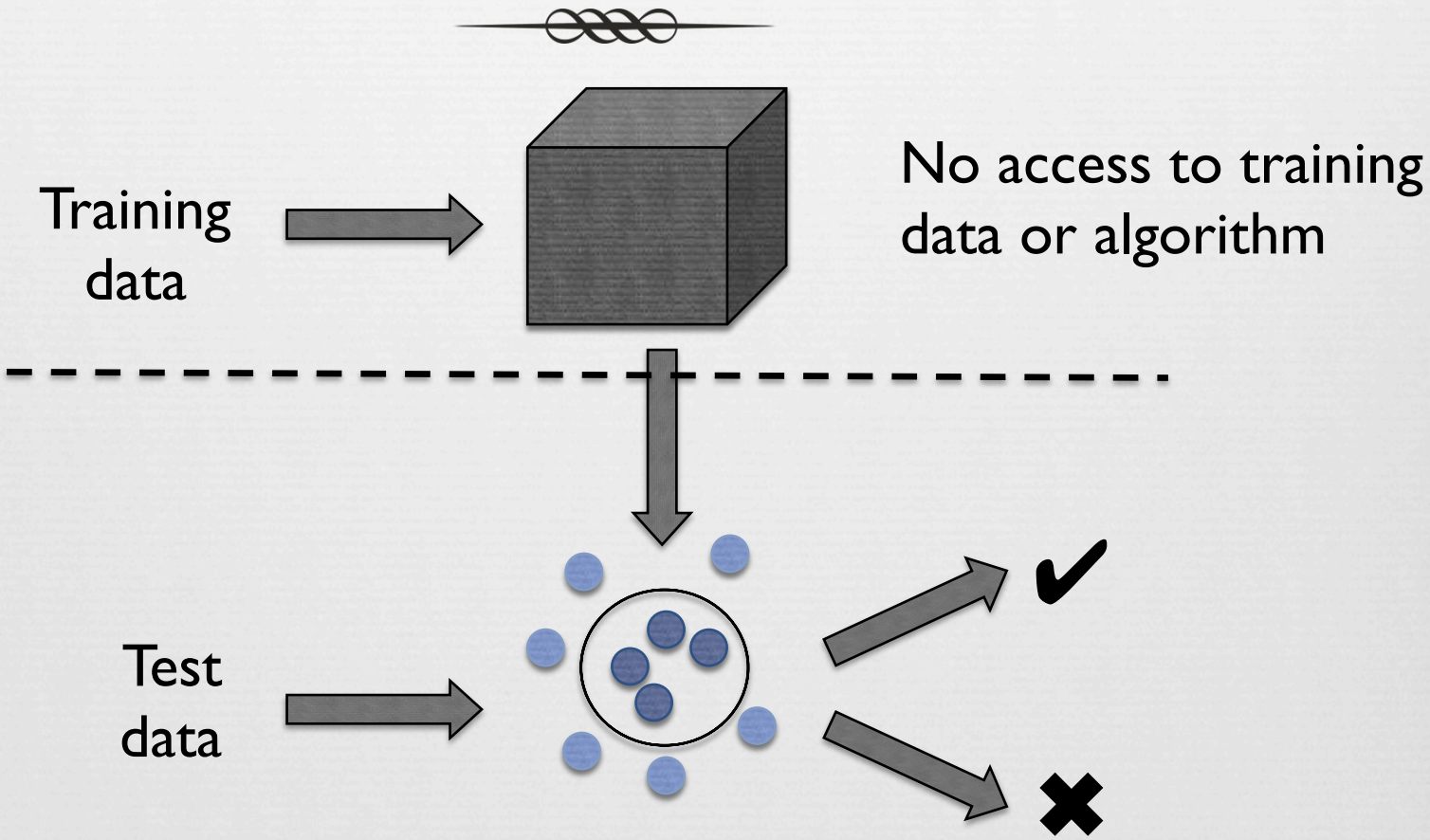
Carlos Scheidegger
Suresh Venkatasubramanian
Charles Marx

Make sure we can query black box algorithms

In six major same-day delivery cities, however, the service area excludes predominantly black ZIP codes to varying degrees, according to a Bloomberg analysis that compared Amazon same-day delivery areas with U.S. Census Bureau data.

http://www.bloomberg.com/graphics/2016-amazon-same-day/

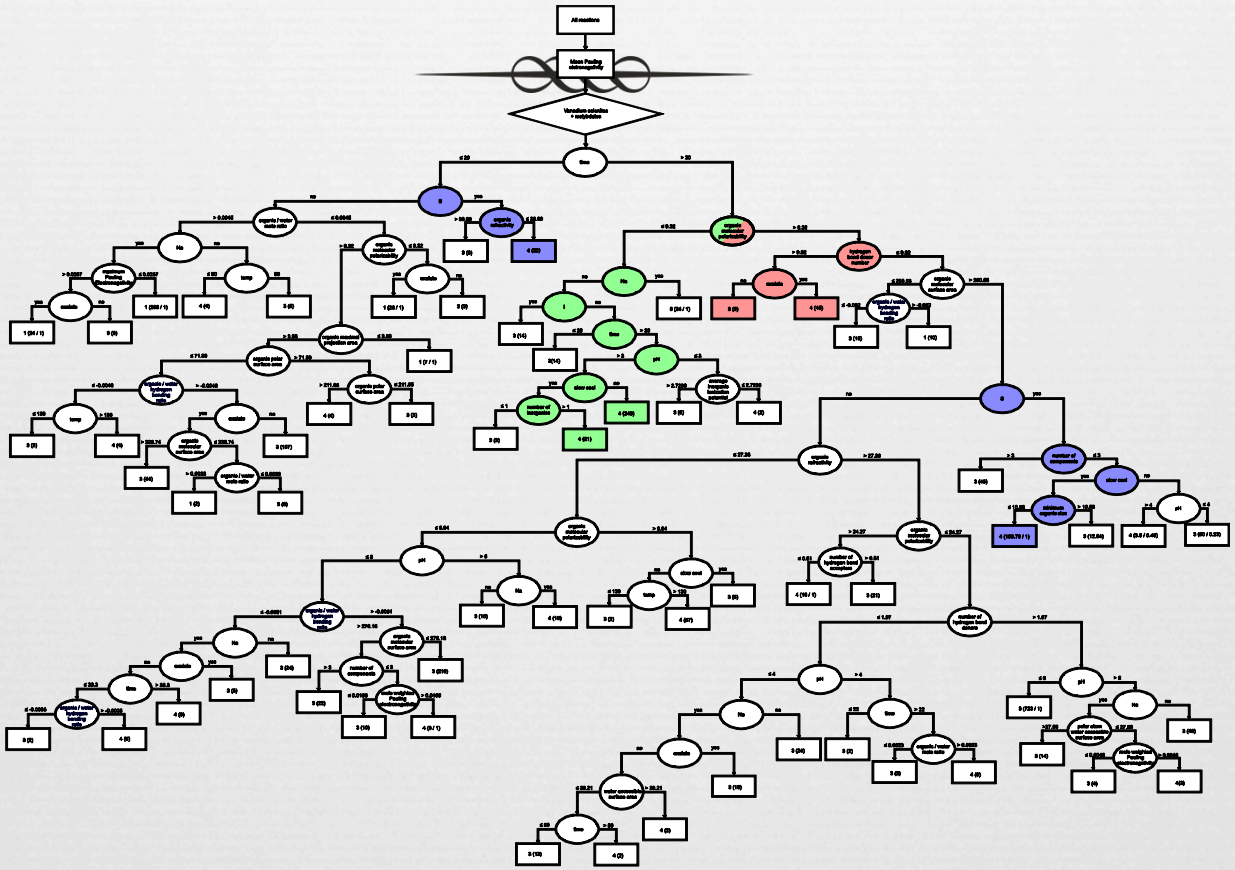
Training vs Testing



How can we understand a model

- If we use a “simple” model we can interpret it directly.
  - Decision trees
  - Linear classifiers
  - SLIM (Sparse Linear Interpretable Models)

Simple models are hard



Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. Nature, 533: 73 - 76, May 5, 2016. http://dx.doi.org/10.1038/nature17439

Research Question

- Given a black box function

$$Y = f(x_1, \ldots, x_n)$$

- Determine the *influence* each variable has on the outcome
  - How do we quantify influence
  - How do we model it (random perturbations?)
  - How do we handle *indirect* and *joint* influence

Direct vs Indirect Influence Auditing

- Does a feature (or group of features) **directly** influence the outcome?
  - E.g a feature used in a decision tree
- **Intervention:**
  - Replace feature with random noise and see how much model accuracy degrades.

Direct vs Indirect Influence Auditing

- Does a feature (or group of features) in**directly** influence the outcome?
  - E.g zipcode as a proxy for race?
- **Intervention:**
  - Direct perturbation no longer works, because more than one variable carries the desired signal.

Information content and indirect influence

the information content of a feature can be estimated by trying to predict it from the remaining features

If the removed feature can't be predicted from the remaining features, then the information from that feature can't influence the outcome of the model.

Information content and indirect influence

the information content of a feature can be estimated by trying to predict it from the remaining features

Given variables X, Y that are correlated, find Y' conditionally **independent** of X such that Y' is as similar to X as possible.

Gradient Feature Audit

For each feature,

1. Remove indirect influence of feature on other features in data
2. Run **model** on modified test data
3. Feature influence = original accuracy – resulting accuracy

Example: Auditing Amazon model:

Feature to remove: *race*

Eliminate (obscure) influence of race on *zipcode*

Gradient Feature Audit

For each feature,

1. Remove indirect influence of feature on other features in data
2. Run **model** on
3. Feature inf ng accuracy

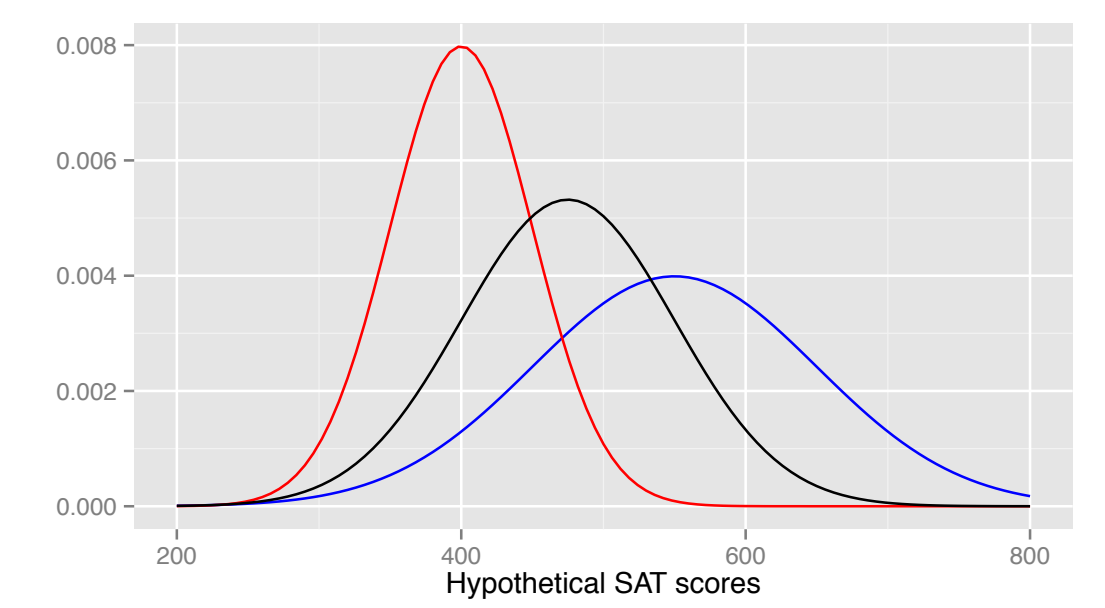
Example: Auditin

Feature to

Eliminate (pcode

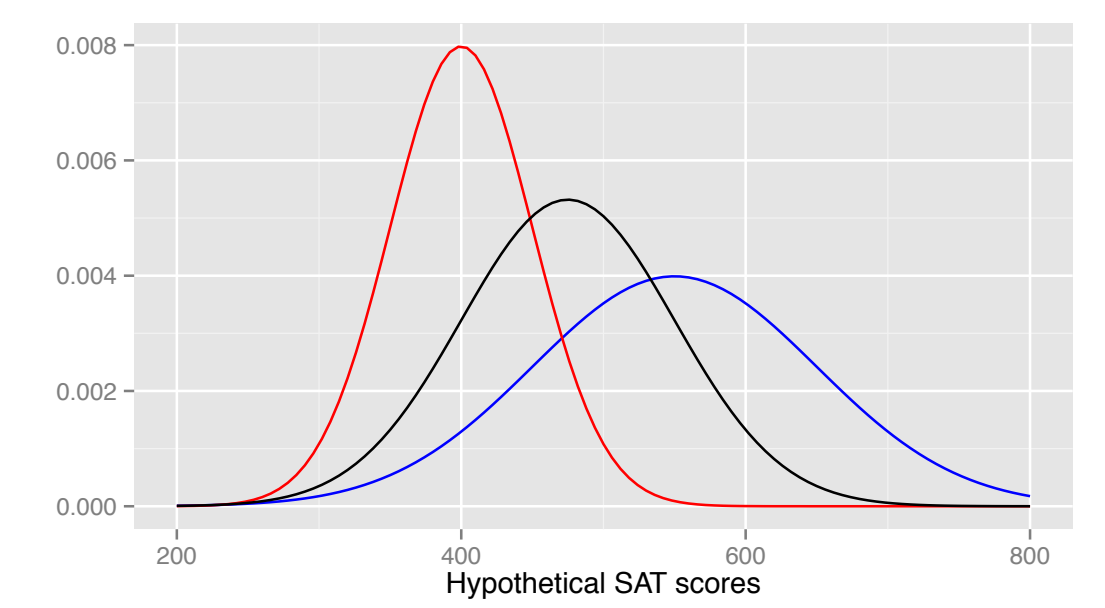
All our measures of influence are relative to a fixed model.

How do we remove indirect influence?



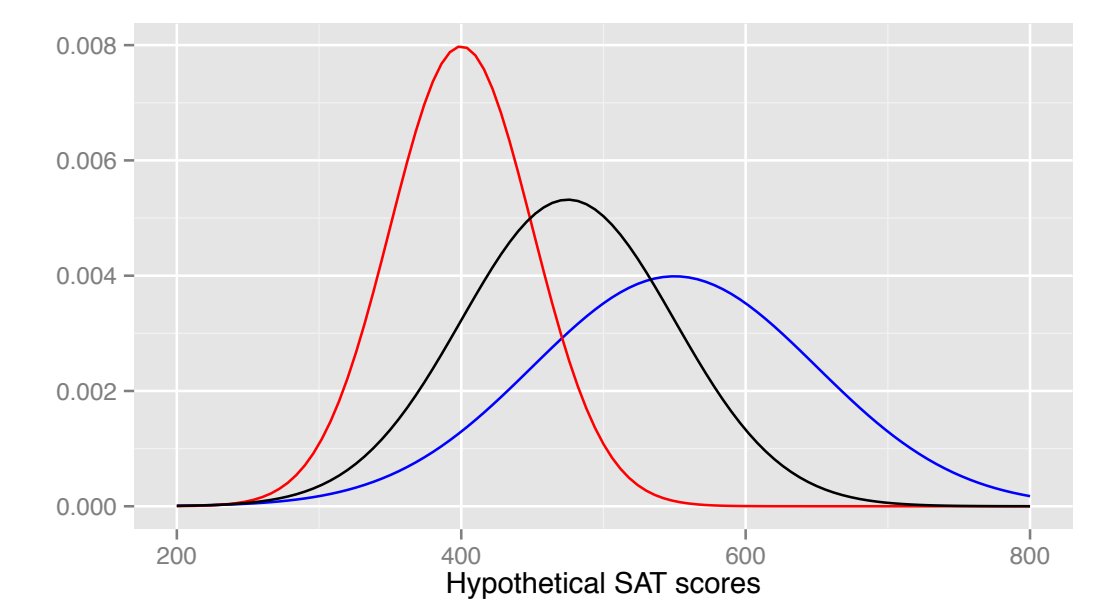
Merge conditional distributions of obscured feature based on eliminated feature.

How do we remove indirect influence?



This will ensure that F-test will fail to tell them apart (provably*)

How do we remove indirect influence?



Need different approaches for categorical and numerical removed and eliminated variables.

Representation matters!

- Should race be categorical or numerical?
- Should it be “white/non-white” or multi-valued?
- These issues matter! For more, see
  - https://arxiv.org/abs/1802.04422
  - https://github.com/algofairness/fairness-comparison