Prediction, Causation, and Interpretation in Social Science

Duncan Watts
Microsoft Research

Explanation in Social Science: Causation or Interpretation?

- When social scientists talk about "explanation" they (almost always) mean "causal explanation"
  - What would have happened to Y had X been different? (Woodward, 2005)
  - Standard counterfactual model of causal inference (Rubin, 1974)
- Yet when social scientists evaluate explanations, causal validity is often not (properly) established
  - Causal inference is hard, and required assumptions are rarely satisfied in practice (Manski 2007; Morgan and Winship 2007; Dunning 2012)
  - Counterfactual model only applies to "effects of causes;" yet many "why" questions are "causes of effects" (Gelman and Imbens, 2013)
- Instead, explanatory power is typically evaluated in terms of plausibility
  - True for quantitative as well as qualitative work (e.g. hypothesis testing makes "predictions" about the sign and significance of coefficients (β) not the outcomes (*y*); model accuracy is rarely emphasized, yet findings are presented as "explanatory" or "predictive" anyway (Yarkoni and Westfall, 2017))

Conflation of Two Meanings

- Conflation of two meanings of "explanation" goes back to Weber:
  - sociology "is a science concerning itself with the **interpretive understanding** of social action and thereby with a **causal explanation** of its course and consequences" (Weber, 1968)
- But they are fundamentally different:
  - Causal explanations necessarily make predictions (although predictive accuracy is not sufficient for casual validity), hence
    - Must be specified ex-ante (or at least in presence of held-out data)
    - Are evaluated in terms of accuracy
  - Interpretively satisfying explanations need only "make sense" of known outcomes, hence
    - Can be (and usually are) specified ex-post
    - Are evaluated in terms of plausibility
- Neither meaning necessarily implies the other
  - Possible to have interpretable explanations that are not predictively accurate and vice versa

Red Herrings

- Social scientists often react negatively to calls to improve predictive accuracy
  - “Predictions do not imply causality”
  - “Complex models generalize poorly”
  - “Uninterpretable models do not provide insight, hence do not aid scientific understanding”
- These are all red herrings
  - Predictions do not imply causality but causality does imply predictions (Hempel and Oppenheim, 1948; Manski, 2007, etc.)
  - Complex models may or may not generalize worse than simple models (Domingos, 1999)
  - Insight is neither causality nor generalization, hence relationship to scientific understanding is unclear

Being Clear

- Ultimately there are two reasons to care about interpretability
  1. For its own sake (i.e. because intuitive understanding is intrinsically rewarding)
  2. Because it is a proxy for generalization/causation
- These reasons have very different validity
  - (1) is a subjective preference so cannot be wrong
  - (2) is wrong in principle and often wrong in practice
    - If generalization and/or causation matter, better to test for them directly
- Fine to value either, or both, but misleading to offer (2) when in fact (1) is really what is being sought

Interpretation still valuable

- In a perfect world, would have explanations that are both interpretable and accurate
  - Some efforts in ML to achieve this (e.g. model selection, regularization, ex-post approximation)
  - Also many instances where simple models perform indistinguishably from complex models (Dawes, 1979; Goldstein, Goel et al., 2017)
- Even when not an end goal, interpretation can be procedurally useful
  - Specifying hypothesis, building intuition about the data, selecting models, modifying hypotheses, etc.

Parallels With Decision Systems

Social Scientists

- Seek explanations of social processes and outcomes
- Want explanations to be interpretable and valid
- Interpretability valued for its own sake
  - Subjective preference
- Interpretability as a proxy for causation/generalization
  - Neither necessary nor sufficient
- Conclude that interpretability is not substitute for validity
  - Validity should be established directly
- But interpretation still valuable
  - For building intuition, refining hypotheses, etc.

Decision Systems

- Seek to make decisions about people
- Want decisions to be understandable and accurate
- Explanation valued for its own sake
  - Subjective preference
- Explanation as a proxy for accountability
  - Neither necessary nor sufficient
- Conclude that explanations are not a substitute for accountability
  - Accountability should be established directly
- But explanations still valuable
  - For exploring systems, refining questions, etc.