MY474: Applied Machine Learning for Social Science

Lecture 2: To Explain or Predict

Blake Miller

09 October 2019

Concepts

1. Explain vs. predict
2. Big data
3. Fundamental problem of causal inference
4. Exploratory data analysis, hypothesis generation

To explain or predict (β vs. $\hat{y}$)

- Can use the same model to explain and predict, but the approaches are distinct!
- Explain
  - Causality comes from theory, usually not the model itself
  - Inference, parameter estimates help us explain causal relationships
  - Simpler model, variables that are statistically significant, not collinear
  - Model should be close to "true model"
  - Look at p-values
- Predict
  - Predict future outcomes given past observations
  - Evaluated based on generalization error
  - Collinearity does not matter
  - Model does not have to resemble a "true model"

"Big Data" and Machine Learning (think, pair, share)

1. What is "big data?" What does it have to do with machine learning?
2. How might "big data" aid in social science research?
3. What are the limitations of big data approaches?
4. What must we be cautious about when using "big data?"

"Big Data" and Machine Learning

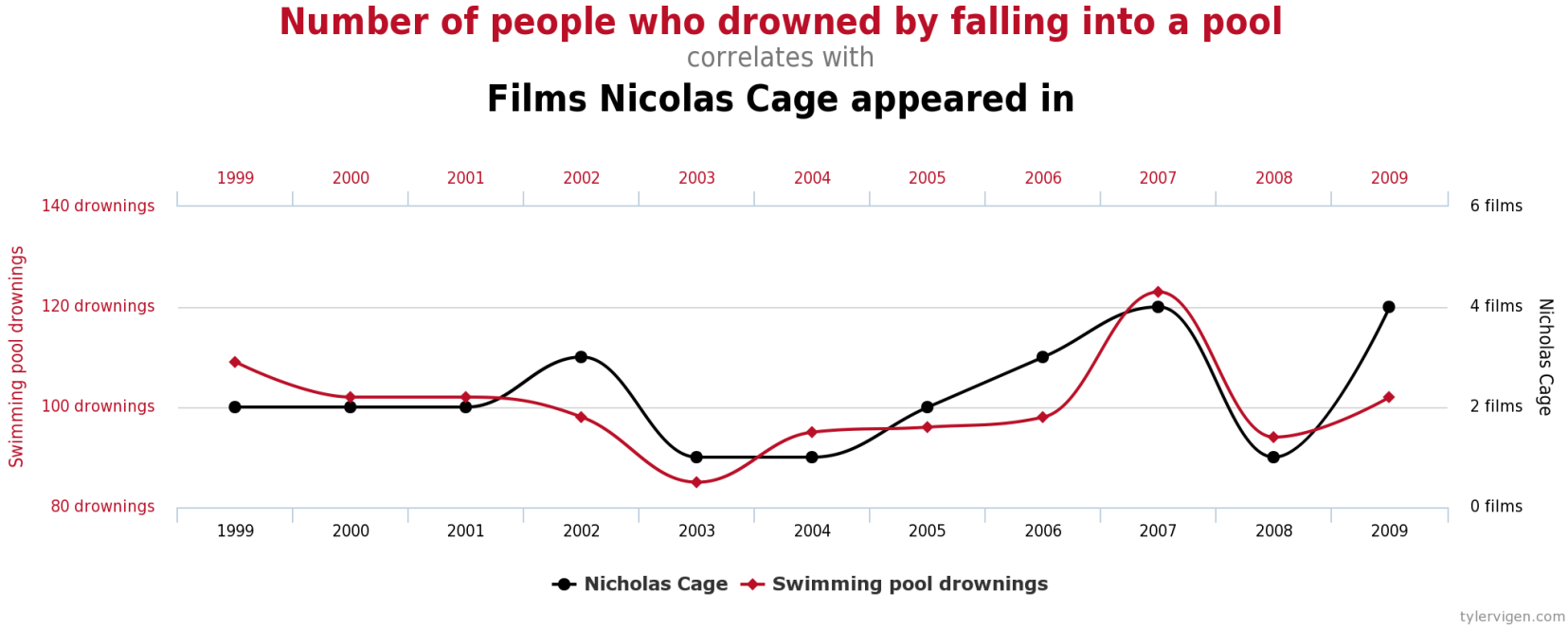
- Machine learning allows us to model, explore, and analyze massive datasets in ways that traditional statistical techniques cannot
- Big data is often in text form, and encodes in it social phenomena we wish to study
- Big data and machine learning for social science research must be paired with a careful research design

Correlation vs. Causation

Usually, as social scientists, we are interested in a causal relationship:

$$X \overset{causes}{\rightarrow} Y$$

Correlation vs. Causation



But correlation, of course, does not imply causation.

Causal Inference

- Treatment, t
- Control, c
- Unit, u
- Potential outcomes $Y_t(u)$ and $Y_c(u)$
- If we could know both outcomes for a single unit, the treatment effect for unit u would be $Y_t(u) - Y_c(u)$

Fundamental Problem of Causal Inference

- It is impossible to observe both $Y_t(u)$ and $Y_c(u)$, so we cannot calculate the treatment effect $\delta = Y_t(u) - Y_c(u)$ (Holland 1986)
- In other words we observe one potential outcome and do not observe the **counterfactual**
- Because of this, we are interested in measuring the average causal effect $E[\delta] = E[Y_{ti}] - E[Y_{ci}]$
- Causal inference involves "a search for assumptions under which we can infer the values of these unobserved **counterfactual** outcomes from observed data" (Titiunik, 2015)

Fundamental Problem of Causal Inference (think, pair, share)

Can "big data" solve the fundamental problem of causal inference? Why/why not?

Takeaways

1. Big data cannot make up for poor theory and research design
2. Big data facilitate rich description, exploration, and hypothesis generation
3. Big data can accelerate of the process of hypothesis generation.
4. "As more phenomena become quantifiable, the range of implications of scientific theories that can be tested empirically is expanded significantly."