CAUSAL MODELING VS PREDICTIVE MODELING

Explanatory modeling

- Refers to the application of statistical models to data for testing causal hypotheses and about theoretical constructs
- From previous lecture
- Designed experiments – DAGs – propensity models
- Aim at estimating a causal relationship – not completely explaining variation in the data

Predictive modeling

- Applying statistical models or data mining for the purpose of predicting new or future observations
- Point estimates (single number), interval prediction, predict distributions
- Not interested in the effect of single variables

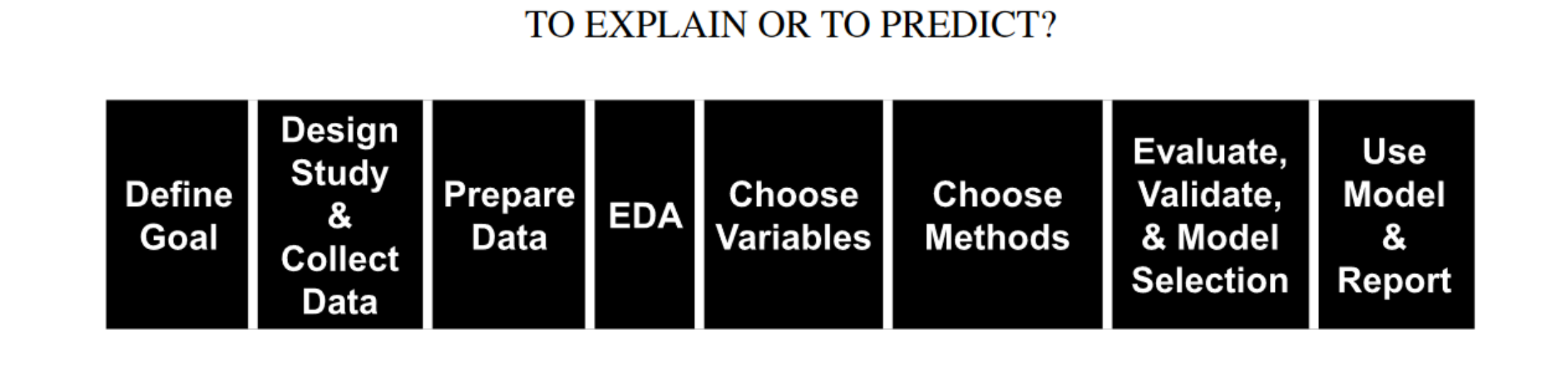
Scientific value of predictive modeling

- Can discover new potential causal mechanisms
- Capture underlying patterns
- Smoothing of data – allows see potential latent variables
- Prediction is related to some causal methods – counterfactual prediction
- LTFU prediction

Differences in explaining and predicting

- Causation – Association
- Theory driven – Data driven
- Retrospective – Prospective
- Bias-Variance

Two modeling paths



Design study and Data collection

Explanatory modeling

- Sample size determined by having precision to estimate the theoretical constructs
- Experimental data preferred
- Reliable and valid instrument that represent underlying construct

Predictive modeling

- Sample size determined by minimizing bias-variance and taking into account potential hold out datasets
- Observational preferred
- Measurement quality and relationship between the variable collected

Data preparation

Explanatory modeling

- Missing data
- Use of imputation
- Data portioning not commonly used – decreases power and precision

Predictive modeling

- Missing data
- Use indicators of missing data
- Data partitioning to address overoptimism

Exploratory data analyis

Explanatory modeling

- Data visualization
- Focuses on exploring data around theoretical question you are asking
- Data reduction PCA used to define underlying constructs

Predictive modeling

- Data visualization
- More freeform, discovering potential relationships not yet known
- Data reduction PCA used to decrease sampling variance

Choice of variables

Explanatory modeling

- Focuses on theoretical relationship
- DAGs
- Previous lectures

Predictive modeling

- Focuses on association with the predictive outcome
- Variables must not only precede outcome but must be available at time of predicition

Choice of methods

Explanatory modeling

- Focuses on interpretability
- Use of regression methods and associated methods

Predictive modeling

- Focuses on predictive accuracy and minimizing bias-variance
- If using regression the coefficients should not be interpreted
- Allows use of data mining, neural networks and shrinkage/penalization methods

Model evaluation and model selection

Explanatory modeling

- Validation
- Focuses on whether model adequately represents/fits the data
- Construct validation – reliability, validity
- Multicollinearity a problem
- Model evaluation assessed by measures that report explanatory power R^2 type values

- Model selection – variables used based on theoretical understanding – see previous lectures

Predictive modeling

- Validation
- Focuses on generalization – ability to predict new observations
- Biggest danger overfitting –focuses on evaluating this
- Multicollinearity less of a problem
- Model evaluation assess by predictive performance

- Model selection – based on predictive performance and may employ a automated selection procedure – stepwise regression
- End-use needs to be taken into account

Shmueli, Galit. “To Explain or to Predict?” Statistical Science 25, no. 3 (August 2010): 289–310. https://doi.org/10.1214/10-STS330