School of Public Health
UNIVERSITY OF CALIFORNIA, BERKELEY

Pre-Specified Analysis Plans and Learning from Data:

Where Are We and Where Are We Going?

Maya Petersen, Alan Hubbard, Mark van der Laan
Div. of Biostatistics, School of Public Health,
University of California, Berkeley

Outline

1. Why have a pre-specified analysis plan?
   - A statistical perspective
2. Current practice in biomedical research
   - FDA and drug efficacy trials
3. Limitations of current practice
   - We need more from our data
4. Pre-specified data adaptive analysis methods
   - Our groups' research program
5. Future directions

The “perfect” study

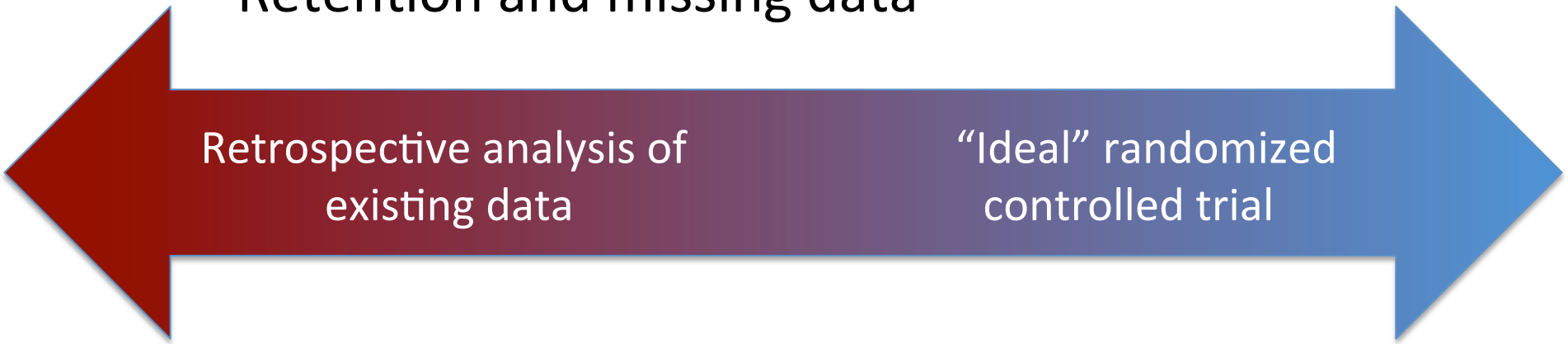
- Able to sample from target population of interest
  - Control the sampling mechanism
- Intervention of interest can be randomized
  - Control the assignment mechanism
- Able to measure the outcome of interest
  - At time point of interest (often long after intervention)
  - No measurement error
- Able to prevent missingness/ losses to follow up

Even in ideal study, many decisions to be made

- Choice of question (target casual quantity)
  - Choice of outcomes
  - Subgroup effects/ effect modification
  - "As treated" (or "Per Protcol")effects
    - Efficacy vs effectiveness in setting of incomplete compliance
  - Direct effects/ effect mediation
    - Ex. Randomized trial of diaphragms showed no impact on HIV infection
    - Is this because any biological effect was masked by decreased condom use?

Real studies are rarely perfect

- Continuum of study designs, with varying degrees of control/knowledge about
  - What population is studied
  - Who is sampled and what data are collected
  - What the intervention is and how it is assigned
  - Compliance
  - Retention and missing data



Even more analytic decisions to be made

- The less we can control in the design phase, the more we rely on analysis
- For a given question, choice of estimator
  - Ex. Estimating the effects of non randomized exposures require some form of adjustment
  - Lots of possible approaches (identification results)
    - Rely on different non-testable assumptions
  - Implementing any one of these involves many decisions
    - What variables to adjust for? How?
  - Analogous decisions required in context of non-random missingness, biased sampling

Statistical Inference

- Goals:
  - Quantify statistical uncertainty
  - Confidence Intervals and hypothesis testing
- Relies on having a well-defined experiment
  - What are the data?
  - How were they sampled, from what population?
  - How were they analyzed? How is estimator defined?
    - An estimator is an algorithm (eg. a computer program)
- We can then think about repeating this experiment many times
  - Bias: how far is mean estimate from the truth
  - Variance: how much does the estimate vary across repetitions

The dangers of approaching our analysis *ad hoc*

- If we do not have a pre-specified analysis plan, we no longer have a well-defined experiment
- What is the experiment?
  - We can think of resampling data in same way from same population…
  - But then, each time we submit it to a (different?) panel of experts?
  - Run some regressions, confer, decide which results make most sense….?

The dangers of approaching our analysis *ad hoc*

- If this process is not formally specified, how do we incorporate it in our inference procedures?
- If we ignore this process

1. Misleading (under) estimate of uncertainty
2. Bias
   - Humans are good at creating narratives from complexity
   - Tendency to confirm what we expect to find

- As long there is "art" in statistics, we will continue to make a lot of wrong inferences

Response in biomedical research

- 2004: Major medical journals require trial registration as condition for publication

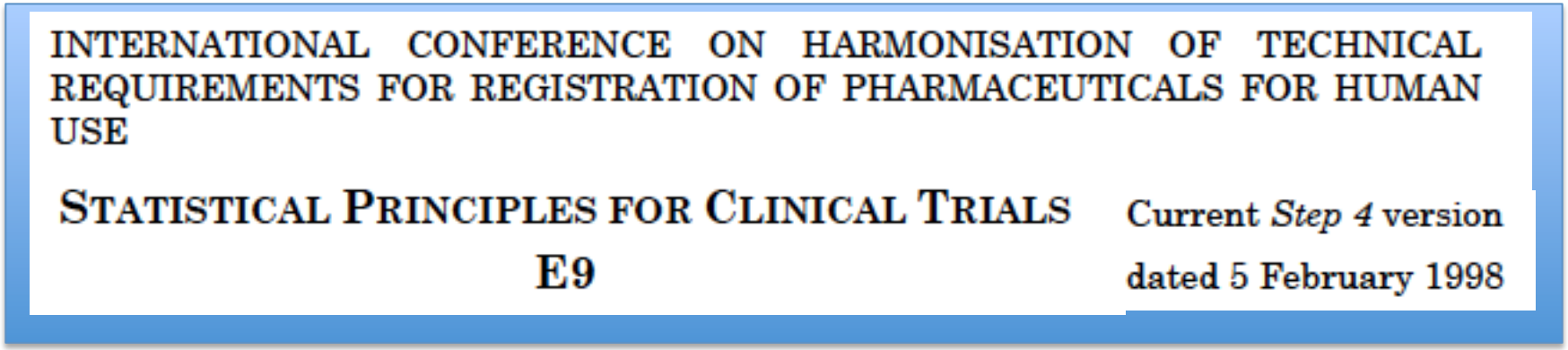
- 2007: Registration of all clinical trials required by US law
  - www.ClinicalTrials.gov
  - Can also register observational studies

Registration of clinical trials

- Registration includes specification of
  - Intervention and comparison
  - Hypotheses
  - Primary and secondary outcome measures
  - Eligibility criteria
  - Key dates
  - Target sample size
  - Funding source
  - Contact information for PI
  - Results (reporting requirements expanded 2007)

Pre-specified analysis plans in clinical trials

- Phase II and III drug trials must submit formal trial protocol prior to start
  - Includes analysis plan



INTERNATIONAL CONFERENCE ON HARMONISATION OF TECHNICAL REQUIREMENTS FOR REGISTRATION OF PHARMACEUTICALS FOR HUMAN USE

STATISTICAL PRINCIPLES FOR CLINICAL TRIALS

E9

Current *Step 4* version

dated 5 February 1998

- ICH: Drug regulatory authorities and the pharmaceutical industry of Europe, Japan and the United States.

Pre-Analysis Plans and the FDA

- Elements of statistical analysis plan
  - Outcomes
  - Interim analyses and stopping rules
  - Subgroup analyses
  - How data will be analyzed in response to
    - Drop-out/Missing data
    - Intent-to-Treat vs. “Per Protocol” Analyses
  - Approach to variance estimation, inference
  - Approaches to multiple testing

A good system, but still work to be done

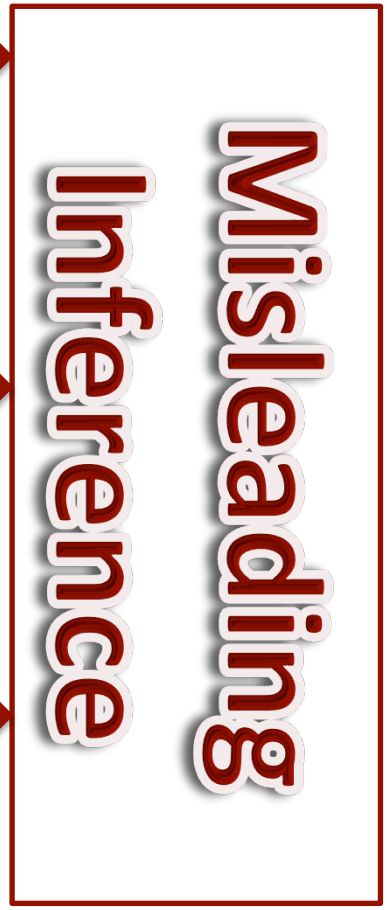
- Once we move beyond simple comparison of randomized groups, existing system does not provide good solutions for addressing
  - Losses to follow up, missing data
  - Per protocol, mediation analysis, effect modification
  - Observational studies
- Why? These are much harder problems
  - Require lots of analytic decisions
  - Not possible/practical to make these decisions without looking at the data…
  - Results can be very sensitive to how you analyze

Example: Study Drop-out

- Common source of bias- drop outs are different
- Goal: Use measured covariates to remove as much bias as possible
- Challenges:
  - We typically measure a lot of covariates
  - Many vary over time and are affected by the exposure
  - Which covariates to adjust for? How?
  - What functional form?
- Generally cannot *a priori* specify a correct parametric model to adjust for bias due to informative drop out

We cannot *a priori* specify a correct parametric model

1. Do nothing?
   – Ignore all that helpful data you collected
2. Use an *a priori* specified parametric model?
   – Even though clear that your model is wrong
3. Look at the data *ad hoc*?
   – Run regressions until things make sense



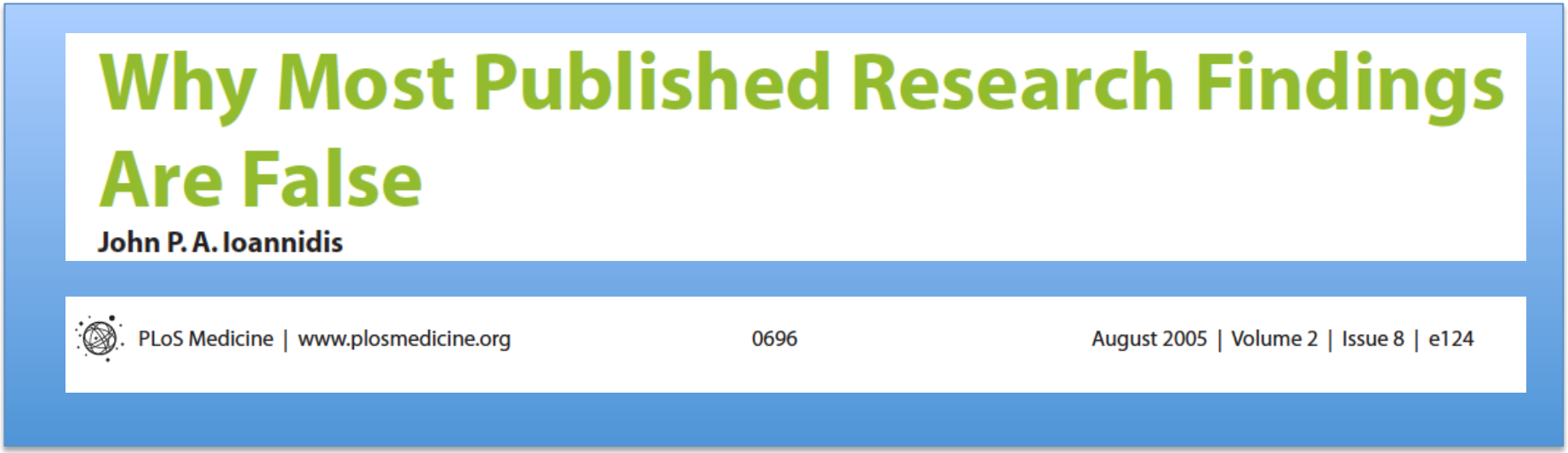
Observational data is even more challenging

- The less we can control in the design phase

1. The more we need an *a priori* analysis plan
   - Results can be very sensitive to analysis decisions
   - Ex. How do you approach confounding
     - Difference in differences versus adjustment for baseline outcome
     - Outcome Regression methods? Propensity score methods?
     - Model specifications?
2. The harder it is to specify such a plan adequately

The Debate: Be careful!

- On the one hand: Growing discomfort with how often we get things wrong



The Debate: Learn more!

- On the other hand: increasing access to huge rich data sets= opportunity
  - Lots of subjects, lots of variables, lots of "complexity" (ie messiness)

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

70 Harvard Business Review **October 2012**

Registration of observational research: The debate continues…

- Clinical journals tend to favor registration

- Epidemiology journals tend to oppose it

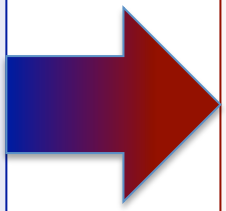
Our research agenda: How to "Have your cake and eat it too"

- Learn more…
  - Use flexible estimators that can respond to the data
  - Data-adaptive or machine learning methods are not just for exploratory analysis
  - The problems we face are hard – if we don't do this we will not get good answers
- But learn rigorously…
  - The estimator is an *a priori* specified algorithm
  - The algorithm itself is flexible
- Objective: Minimize bias, optimize validity of statistical inference

Ultimate Objective:

User Input

- Question
  - Prediction versus causal
  - Point, longitudinal, static, dynamic, stochastic exposures
- Data
  - Longitudinal, Hierarchical
  - Missing data
- Model
  - Causal and statistical
  - Knowledge about data generating process



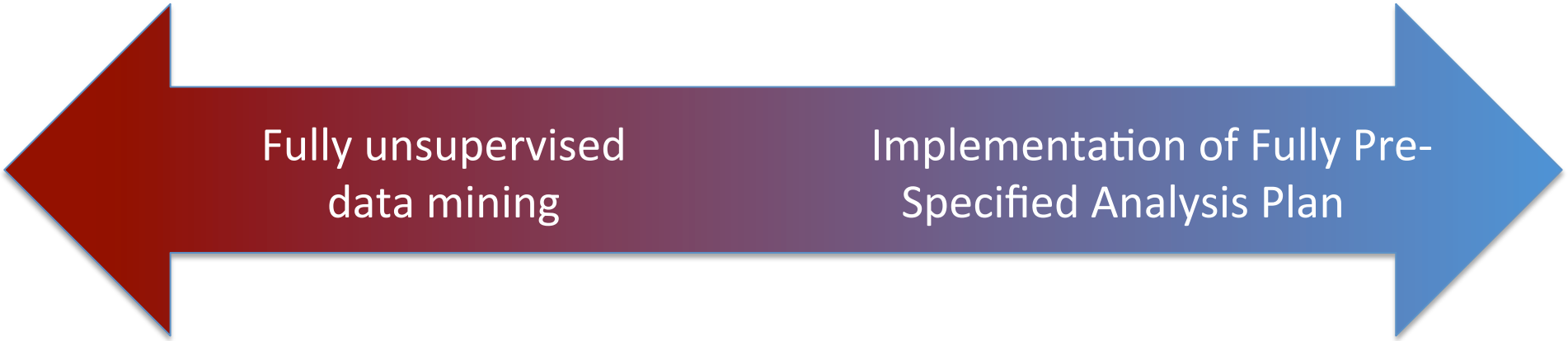
Output

- Target statistical parameter
- Point estimate
- Statistical Inference
- Diagnostics
  - Suggested responses if insufficient support
- Guidance for interpretation
  - Assumptions needed for causal interpretation

Must one always have a PAP?

- There will always be interesting questions that we fail to anticipate
- Our perspective

1. Have a PAP when starting a project
2. Be transparent about deviations
3. Again- there is a continuum!



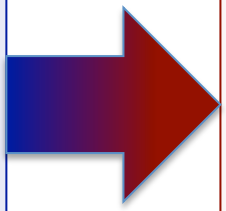
Deviation from PAP does not have to mean abandoning rigor

- Example: Safety Analysis
  - Adverse effect of drugs may be rare of take years to develop
  - Recognizing these requires analysis of large observational datasets
  - Either extreme of the spectrum is risky
- Remains a major role for data adaptive software
  - If you add a non-prespecified question, can still take the "art" out of answering it
  - Even if your target parameter was not *a priori* specified, your estimator should be

So where does the "art" come in???

User Input

- Question
  - Prediction versus causal
  - Point, longitudinal, static, dynamic, stochastic exposures
- Data
  - Longitudinal, Hierarchical
  - Missing data
- Model
  - Causal and statistical
  - Knowledge about data generating process



Output

- Target statistical parameter
- Point estimate
- Statistical Inference
- Diagnostics
  - Suggested responses if insufficient support
- Guidance for interpretation
  - Assumptions needed for causal interpretation

So where does the "art" come in???

User Input

- Question
  - Prediction versus causal
  - Point, longitudinal, static, dynamic, stochastic exposures
- Data
  - Longitudinal, Hierarchical
  - Missing data
- Model
  - Causal and statistical
  - Knowledge about data generating process

- Understanding and articulating the relevant questions
- Understanding the data
- Understanding the experiment that generated it
  - Study design
  - Expert knowledge

How close are we?
Causal Inference Frameworks

- Formal causal frameworks
  - Judea Pearl: Non-Parametric Structural Equation Models/ Causal Graphs
  - Neyman-Rubin Potential Outcome Model
- Minimize unsubstantiated assumptions about causal relationships
  - Ex. Assumptions on functional form, error distributions
- Provide a flexible tool for
  - Rigorously representing background knowledge
  - Translating a wide range of questions into formal queries
  - Determining whether Data + Knowledge are sufficient
  - Determining what additional Data+ Assumptions needed

How close are we? Statistical Theory

- These are hard statistical problems
  - High dimensional data
  - Complex target parameters
- Advances in Machine Learning
  - Ensemble Loss based Learning and Cross-Validation
  - Great for prediction, but not sufficient for casual questions
- Advances in semi-parametric efficient estimation
  - Targeted bias reduction for question of interest
  - Targeted Maximum Likelihood Estimation (TMLE)

How close are we?
Software (Public R packages)

1. Super Learner: `SuperLearner()`
   - Ensemble Machine Learning for Prediction
2. Targeted Maximum Likelihood Estimation: `tmle()`
   - Effect estimation of point treatment exposures
   - Super Learner + targeting for effect parameter
3. TMLE for Longitudinal Data: `ltmle()`
   - Effects of cumulative exposures
   - Dynamic Interventions
   - Censoring, Missing Data
   - To be released 2013

Current and Future work: Adaptive pre-specified analysis plans

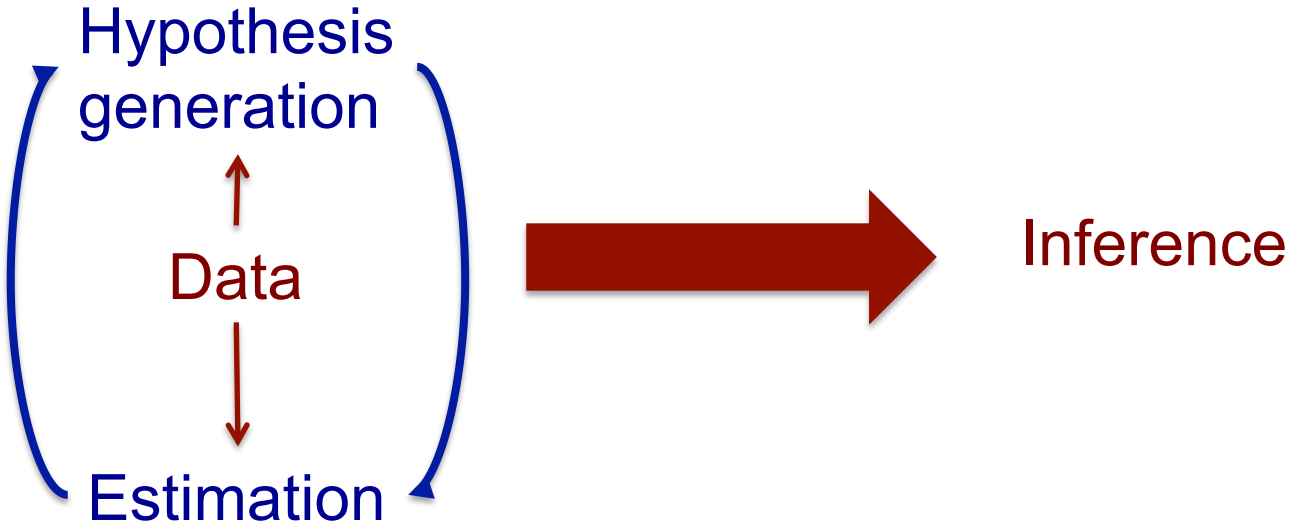
- Changing the question in response to the data
  - In *a priori* specified way: Maintain valid inference

1. Using planned interim analyses of randomized trial to
   - Modify the target population
   - Change randomization probabilities
   - Change the intervention

Current and Future work: Adaptive pre-specified analysis plans

- Changing the question in response to the data
  - In *a priori* specified way: Maintain valid inference

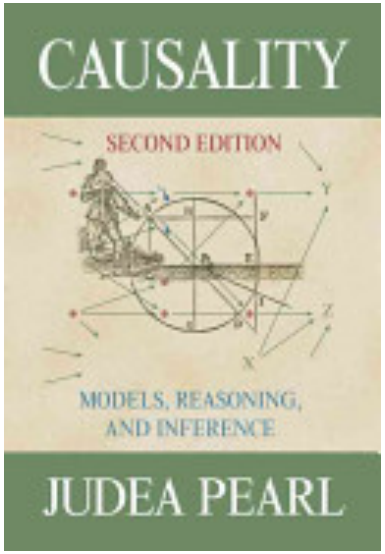
2. *A priori* algorithm for hypothesis selection
   - For a given database, algorithm suggests hypotheses (target parameters) based on data

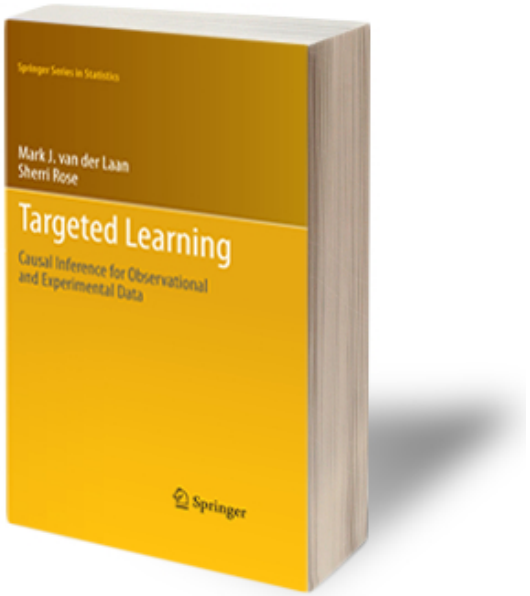


Summary: Our perspective

1. Ideally, *a priori* specify questions
   - Ie. target causal parameters
   - Acknowledge difficulty of specifying all interesting questions *a priori*
2. Regardless of whether question was *a priori* specified, use an *a priori* specified estimator
   - Publically available tools for *a priori* specified data adaptive estimators targeted at the question of interest
3. Eventual goal: an *a priori* specified algorithm that selects target parameters in response to background knowledge and data
   - Theoretical and programming work remains to be done

References & Resources





- http://www.targetedlearningbook.com/
- http://cran.r-project.org/web/packages/
- http://www.stat.berkeley.edu/~laan/
- Journal of Causal Inference (2013-)
  - http://www.degruyter.com/view/j/jci