Machine Learning Methods for Prediction in Epidemiology

Sherri Rose

Assistant Professor of Biostatistics

Harvard Medical School
Department of Health Care Policy

rose@hcp.med.harvard.edu
drsherrirose.com

April 23, 2014

R.A. Fisher

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

Albert Einstein

"To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science."

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The New York Times
nytimes.com

September 16, 2007

Do We Really Know What Makes Us Healthy?

By GARY TAUBES

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

AMSTATNEWS
The Membership Magazine of the American Statistical Association

Statistics Ready for a Revolution

1 SEPTEMBER 2010 503 VIEWS 2 COMMENTS

Next Generation of Statisticians Must Build Tools for Massive Data Sets

Mark van der Laan, Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley, and Sherri Rose, PhD candidate at UC Berkeley

The New York Times
nytimes.com

September 16, 2007

Do We Really Know What Makes Us Healthy?

By GARY TAUBES

variations

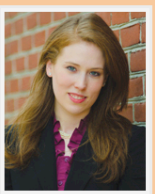
Big data and the future

At the beginning of her career **Sherri Rose** discusses big data and stands amazed at its potential.

STATtr@k A website for new statistics professionals navigating a data-centric

Statisticians' Place in Big Data

– FEBRUARY 1, 2013
POSTED IN: DEVELOPMENT TR@K



Sherri Rose is an NSF mathematical sciences postdoctoral research fellow in the department of biostatistics at the Johns Hopkins Bloomberg School of Public Health.

Big Data has become the new buzz phrase in the world of information collection and analysis. The experiments we conduct and the observational data we collect continue to grow in size, due to rapidly expanding technology.

Large data sets also have drawn the attention of young people, with undergraduate and graduate students choosing computer science, engineering, and statistics for their programs of study. Each of these disciplines brings something unique to the table when discussing the challenges of Big Data, and interdisciplinary collaborations are becoming increasingly common.

Research

Methodological:

- Robust estimation
- Case-control studies
- Causal inference & comparative effectiveness
- Sequential decision theory (e.g., dynamic regimes)
- High-dimensional longitudinal observational data
- Machine learning in prediction and effect estimation

Current Subject Matter Areas:

- Health care systems
- Mental health
- Chronic disease

Image credit: Josh Lee, @wtrsld

statistician

tell me something interesting (fast)

insert data

big data system

Idreos/Rose

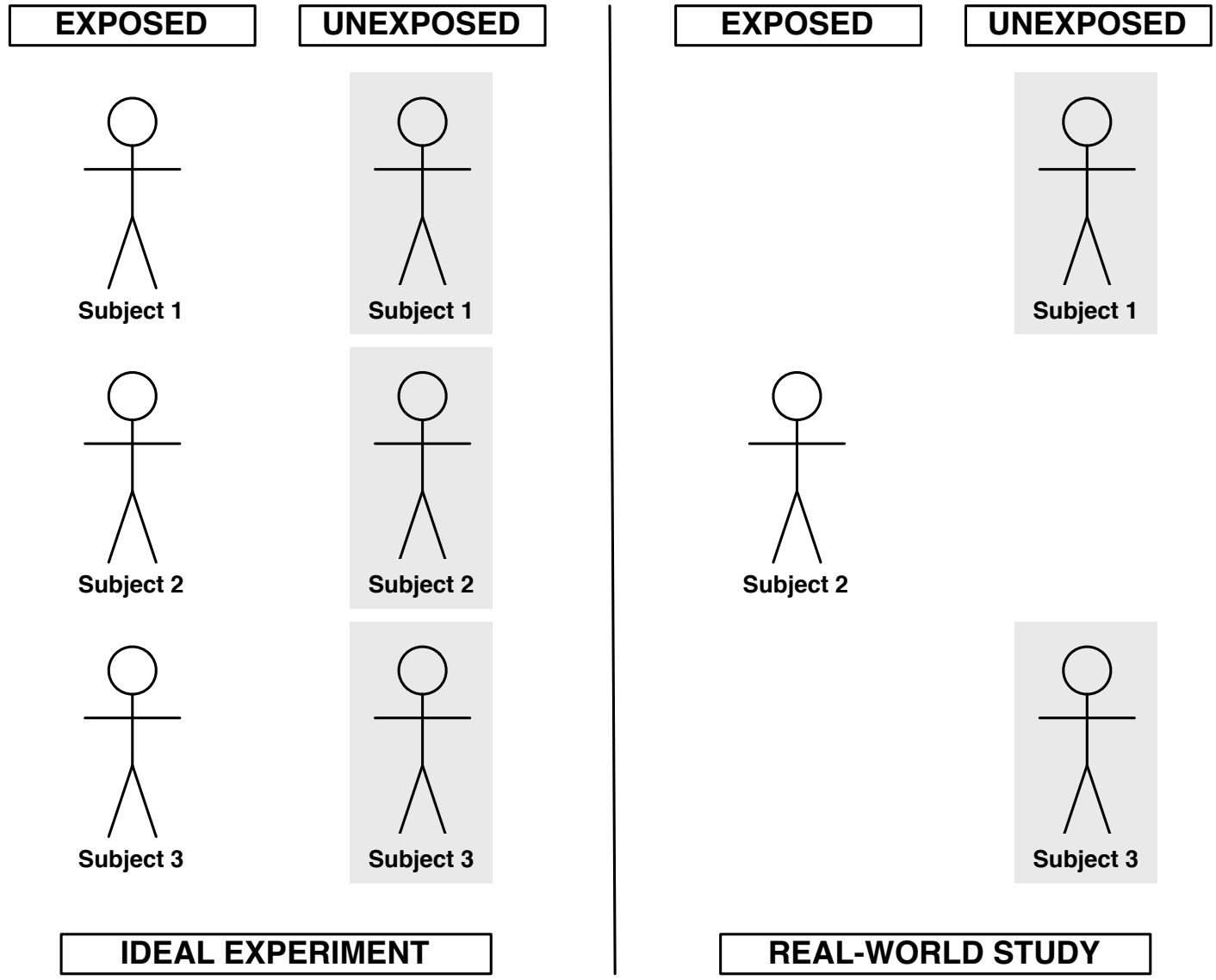
FRAMINGHAM HEART STUDY

A Project of the National Heart, Lung and Blood Institute and Boston University

Breast Cancer Risk Assessment Tool

An interactive tool to help estimate a woman's risk of developing breast cancer

- Austin et al. **Logistic regression had superior performance** compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *JCE.* 2010.
- Peng et al. **Random forest can predict 30-day mortality** of spontaneous intracerebral hemorrhage **with remarkable discrimination.** *EJN.* 2010.


EXPOSED
UNEXPOSED
Subject 1
Subject 1
Subject 2
Subject 2
Subject 3
Subject 3
IDEAL EXPERIMENT
EXPOSED
UNEXPOSED
Subject 1
Subject 2
Subject 3
REAL-WORLD STUDY

Effect Estimation vs. Prediction

Both **effect** and **prediction** research questions are inherently *estimation* questions, but they are distinct in their goals.

Effect Estimation vs. Prediction

Both **effect** and **prediction** research questions are inherently *estimation* questions, but they are distinct in their goals.

Effect: Interested in estimating the effect of exposure on outcome adjusted for covariates.

Effect Estimation vs. Prediction

Both **effect** and **prediction** research questions are inherently *estimation* questions, but they are distinct in their goals.

Effect: Interested in estimating the effect of exposure on outcome adjusted for covariates.

Prediction: Interested in generating a function to input covariates and predict a value for the outcome.

The goal

Want automated algorithm to semiparametrically estimate $E_0(Y \mid W)$.

There are semiparametric methods that also aim to "smooth" the data and estimate this regression function.

Background

Previous studies of elderly populations in the United States have indicated that

- gender
- smoking status
- heart health
- physical activity
- education level
- income
- weight

are among the important predictors of mortality in elderly populations.

Electronic Health Record Databases

The increasing availability of electronic medical records offers a **new resource to public health researchers**.

General usefulness of this type of data to answer targeted scientific research questions is an open question.

Need **novel statistical methods** that have desirable statistical properties while remaining computationally feasible.

Kaiser Permanente Electronic Health Record Database

Kaiser Permanente is based in Northern California and provides medical services to approximately 350,000 persons over the age of 65 each year.

- **Gender** & **age** obtained from administrative databases
- **184 disease and diagnoses variables (medical flags)** obtained from clinical and claims databases

KAISER PERMANENTE®

Kaiser Permanente Electronic Health Record Database

Nested case-control sample (n=27,012) from a Kaiser Permanente database of 345,191 persons over the age of 65 in 2003.

- **Outcome** Y was **death** the subsequent year (2004).
- **Covariates** $W = \{W_1, \ldots W_{186}\}$ were **184 medical flags**, **gender** & **age**.

Observed data structure on a subject can be represented as $O = (Y, \Delta, \Delta X)$, where $X = (W, Y)$ is the full data structure, and Δ denotes the indicator of inclusion in the second-stage sample.

Formalizing the Parameter of Interest

We define our parameter of interest, $Q_0 = E_0(Y \mid W)$, as the minimizer of the expected squared error loss:

$$Q_0 = \arg\min_Q E_0 L(O, Q),$$

where $L(O, Q) = (Y - Q(A, W))^2$. $E_0 L(O, Q)$, which we want to be small, evaluates the candidate Q, and it is minimized at the optimal choice of Q_0. We refer to expected loss as the risk.

Super Learner (van der Laan, Polley, Hubbard; 2007)

Allows researchers to use multiple algorithms to outperform a single algorithm in nonparametric statistical models.

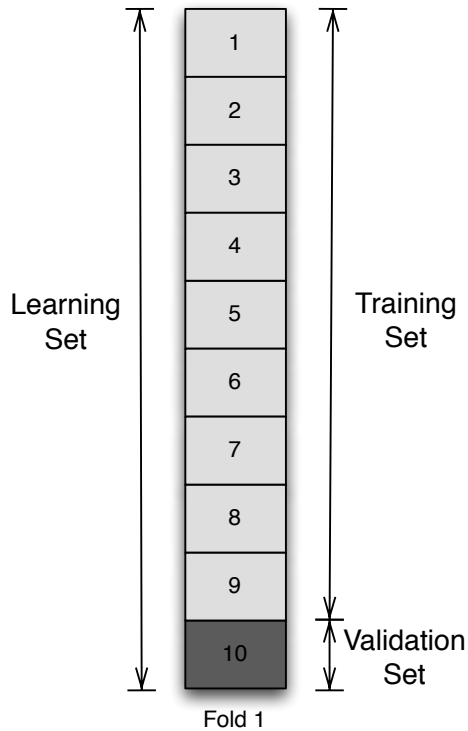
The term algorithm is used very loosely to describe any mapping from the data into a predictor.

Super Learner

- Ensembling methods allow implementation of multiple algorithms.
- Do not need to decide beforehand which single technique to use; can use several by incorporating cross validation.

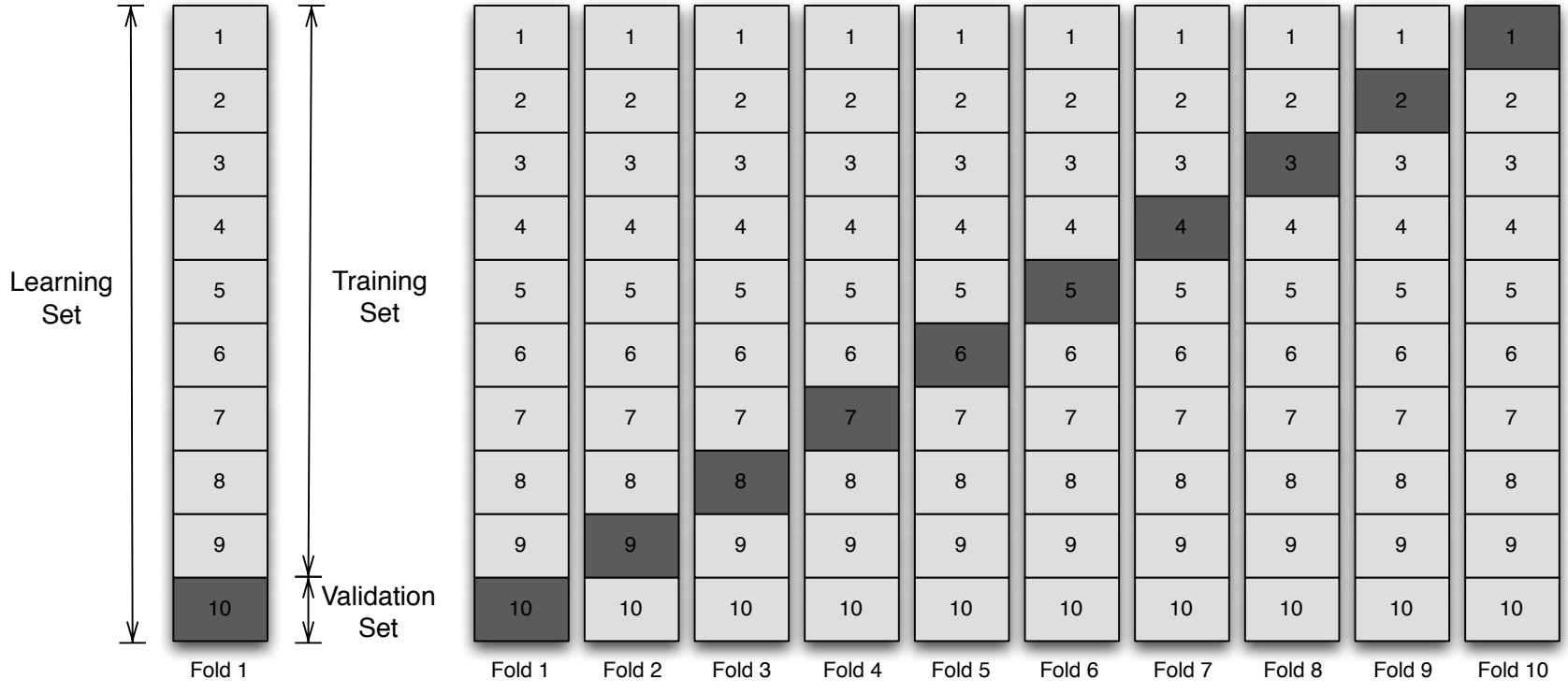
Super Learner

- Ensembling methods allow implementation of multiple algorithms.
- Do not need to decide beforehand which single technique to use; can use several by incorporating **cross-validation**.



Super Learner

- Ensembling methods allow implementation of multiple algorithms.
- Do not need to decide beforehand which single technique to use; can use several by incorporating **cross-validation**.



Super Learner

Build a library of algorithms consisting of all weighted averages of the algorithms.

One of these weighted averages might perform better than one of the algorithms alone.

It is this principle that allows us to map a collection of algorithms into a library of weighted averages of these algorithms.

Super Learner

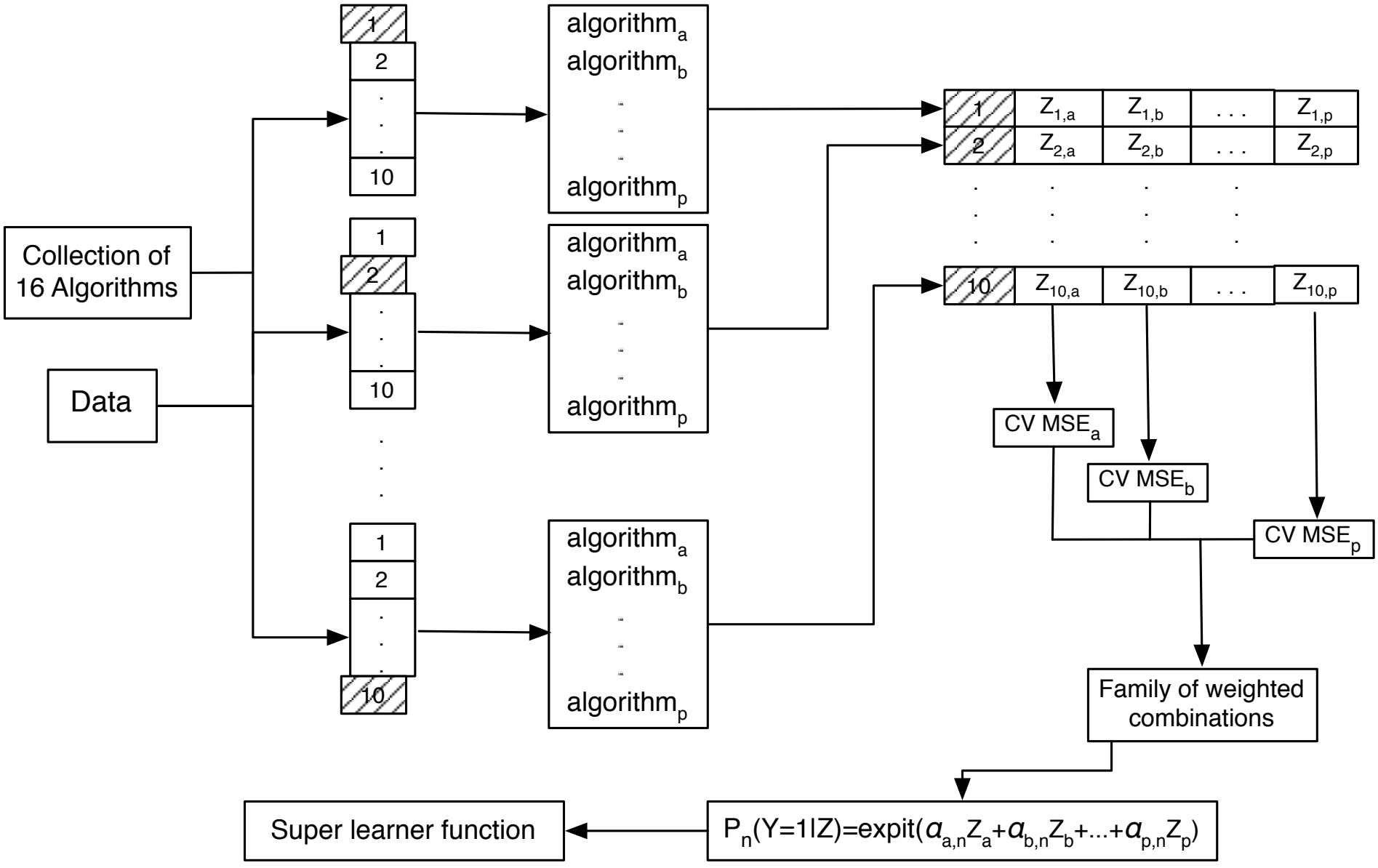
It might seem that the implementation of such an estimator is problematic, since it requires **minimizing the cross-validated risk over an infinite set of candidate algorithms** (the weighted averages).

The contrary is true.

Super learner is not more computer intensive than the "cross-validation selector" (the single algorithm with the smallest cross-validated risk).

- Only the relatively trivial calculation of the optimal weight vector needs to be completed.

Super Learner



More examples[1]

To study the super learner in real data examples, collected a number of publicly available data sets.

- sample sizes ranged from 200 to 654 observations
- number of covariates ranged from 3 to 18
- all 13 data sets have a continuous outcome and no missing values

[1] **Polley, Rose, van der Laan** (2011). Super learning. In: van der Laan, Rose *Targeted Learning: Causal Inference for Observational & Experimental Data.*

Finite sample performance

Table 3.3 Description of data sets, where n is the sample size and p is the number of covariates. All examples have a continuous outcome.

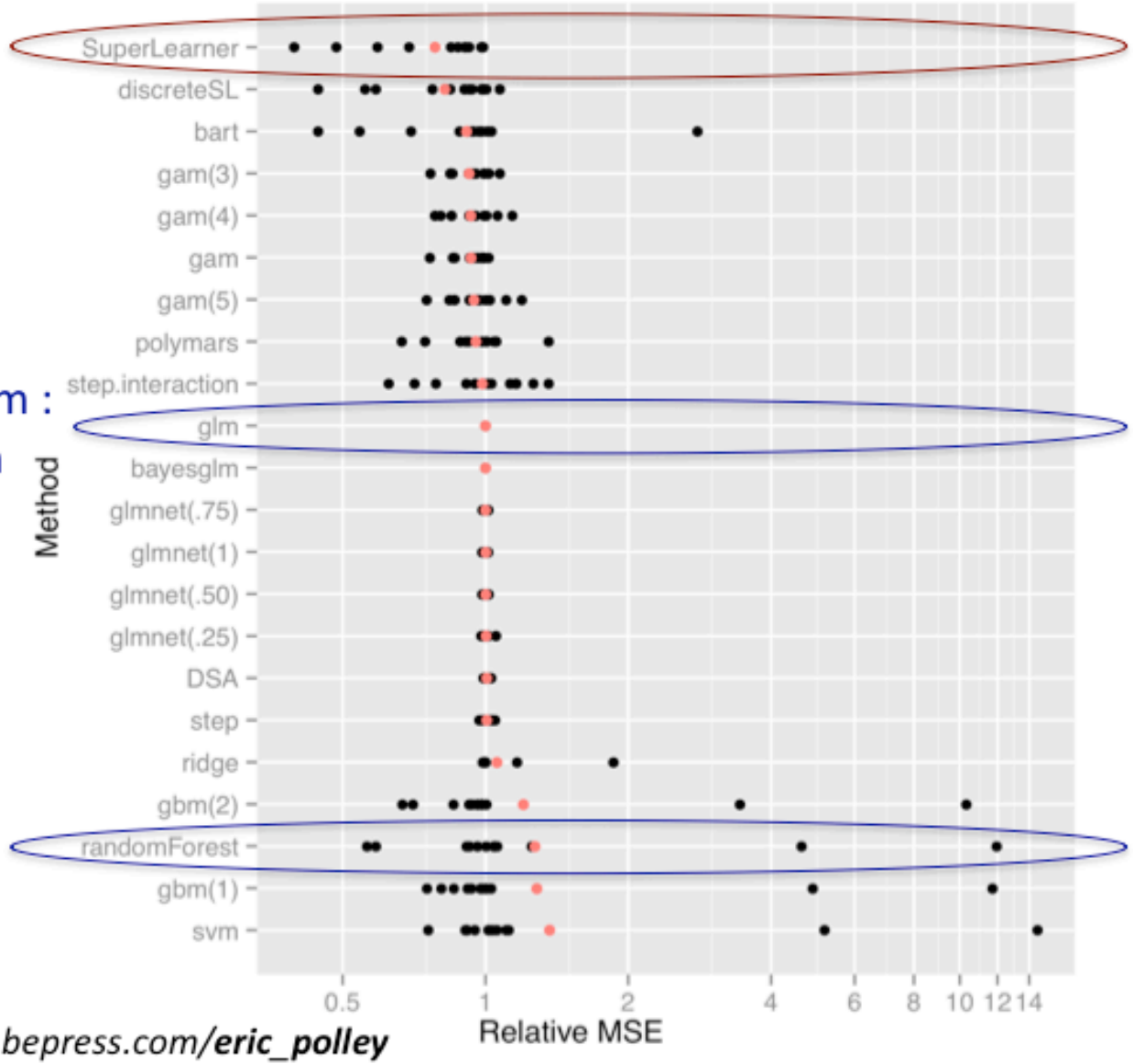
Name	n	p	Source
ais	202	10	Cook and Weisberg (1994)
diamond	308	17	Chu (2001)
cps78	550	18	Berndt (1991)
cps85	534	17	Berndt (1991)
cpu	209	6	Kibler et al. (1989)
FEV	654	4	Rosner (1999)
Pima	392	7	Newman et al. (1998)
laheart	200	10	Afifi and Azen (1979)
mussels	201	3	Cook (1998)
enroll	258	6	Liu and Stengos (1999)
fat	252	14	Penrose et al. (1985)
diabetes	366	15	Harrell (2001)
house	506	13	Newman et al. (1998)

Finite sample performance

Super Learner-
Best weighted combination of algorithms for a given prediction problem

Example algorithm : Linear Main Term Regression

Example algorithm: Random Forest



Technical Report: *works.bepress.com/**eric_polley***

Risk Score Prediction: Kaiser Permanente Database

Ensembling method outperformed all other algorithms.

Generally weak signal with $R^2 = 0.11$.

How will this electronic database perform in comparison to the more traditional cohort study with many fewer subjects and measured covariates?

Risk Score Prediction: Sonoma Cohort Study[2]

Cohort study of $n = 2,066$ residents of Sonoma, CA aged 54 and over.

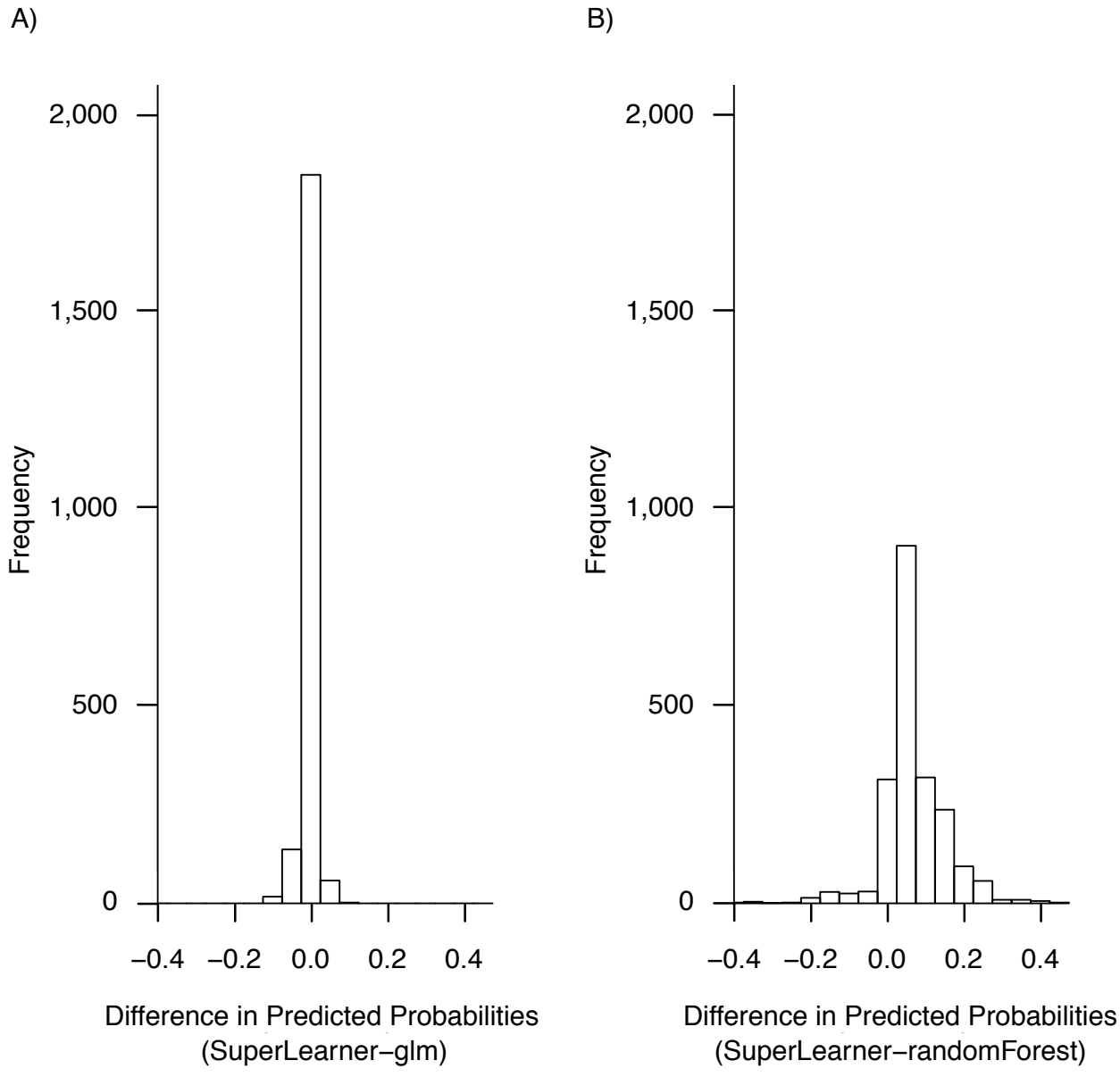
- Outcome was death.
- Covariates were gender, age, **self-rated health**, **leisure-time physical activity**, smoking status, cardiac event history, and chronic health condition status.
- $R^2 = 0.201$

Two-fold improvement occurred with less than 10% of the subjects and less than 10% the number of covariates.

What possible conclusions can we draw from this comparison and the use of electronic medical records in the future?

[2] **S. Rose** (2014). Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*.

Predicted Values



Risk Score Prediction: World Mental Health Study

Survey study with $n = 47,466$ traumatic exposures from around the world.

- Outcome was PTSD.
- Covariates were gender, age at traumatic event, **types of traumatic events**, **prior disorders**, marital status, education, and others.

Our results also based on our final ensembled super learner algorithm placed *[results redacted for public version of slides]* of PTSD outcomes in the top 10% of predicted risk scores

The Need for Targeted Learning in Semiparametric Models

- MLE not targeted for effect parameters.
- Need a subsequent targeted bias-reduction step: Targeted MLE

Targeted Learning

- Avoid reliance on human art and unrealistic parametric models
- Define interesting parameters
- Target the parameter of interest
- Incorporate machine learning
- Statistical inference

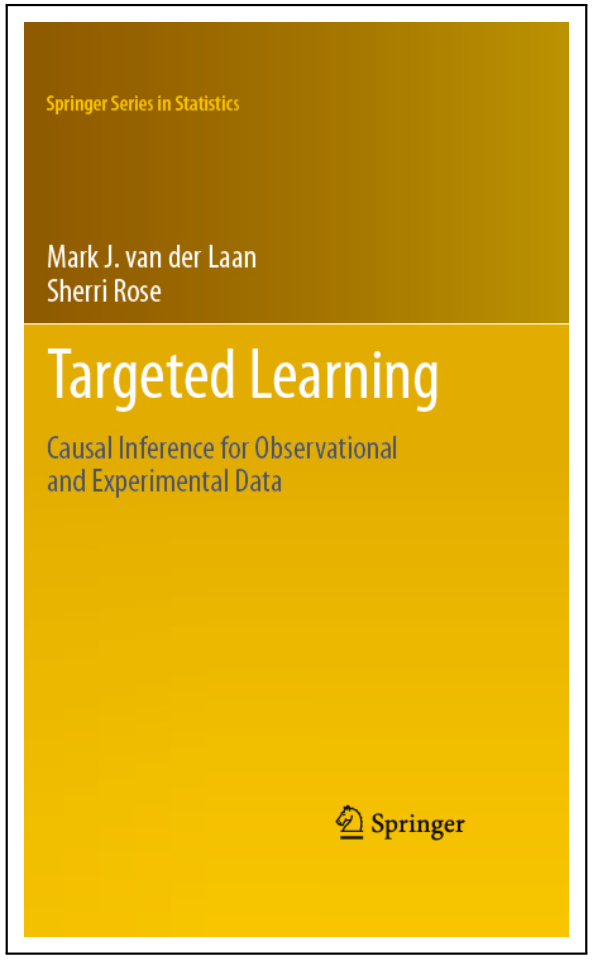
Targeted Maximum Likelihood Learning

Two-step procedure that incorporates estimates of:

1. the probability of the outcome given exposure and covariates
2. the probability of exposure given covariates

With an initial estimate of the outcome regression, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias-variance tradeoff for the parameter of interest.

Targeted Learning Book (targetedlearningbook.com)



van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.