Principles of Prediction and Inference in Machine Learning
Part 3: Performance, Operating Characteristics, and Simulation

Thomas G. Stewart, PhD

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

February 17, 2021

Course goal 5: Distinguish between in-sample and out-of-sample performance and understand the related concept of optimism

Measures of model performance

Prediction

Discrimination

Calibration

Classification

Classification Error

Positive Predictive Value

Negative Predictive Value

Prediction

Discrimination

Are the predictions in the right order?

Calibration

Are the predictions the right value?

Classification

Classification Error

Positive Predictive Value

Negative Predictive Value

Prediction

Discrimination

Calibration

Estimate

Class probabilities

Classification

~~Classification Error~~

~~Positive Predictive Value~~

~~Negative Predictive Value~~

Discrimination

Calibration

AUC / Somers D_{xy} / Concordance

Correlation of Y and $\hat{Y}$ (R^2)

Calibration curve

Maximum absolute calibration error

Brier Score

Aim: Learn something about

Larger population

Future observations

Dataset

Aim: Learn something about

Larger population

Future observations

~~Dataset~~ Not Aim

tension

Dataset

Larger population

Future observations

tension

Dataset

Larger population

Future observations

Discrimination

Calibration

$\neq$

Discrimination

Calibration

tension

Dataset

Larger population

Future observations

In-sample discrimination

In-sample calibration

$\neq$

Out-of-sample discrimination

Out-of-sample calibration

In-sample performance − **Out-of-sample** performance = **Optimism**

Discrimination

Calibration

Optimism corrected
AUC / Somers D_{xy} / Concordance

Optimism corrected
Calibration curve

Optimism corrected
Correlation of Y and $\hat{Y}$ (R^2)

Optimism corrected
Maximum absolute calibration error

Optimism corrected
Brier Score

Course goal 5: Distinguish between in-sample and out-of-sample performance and understand the related concept of optimism

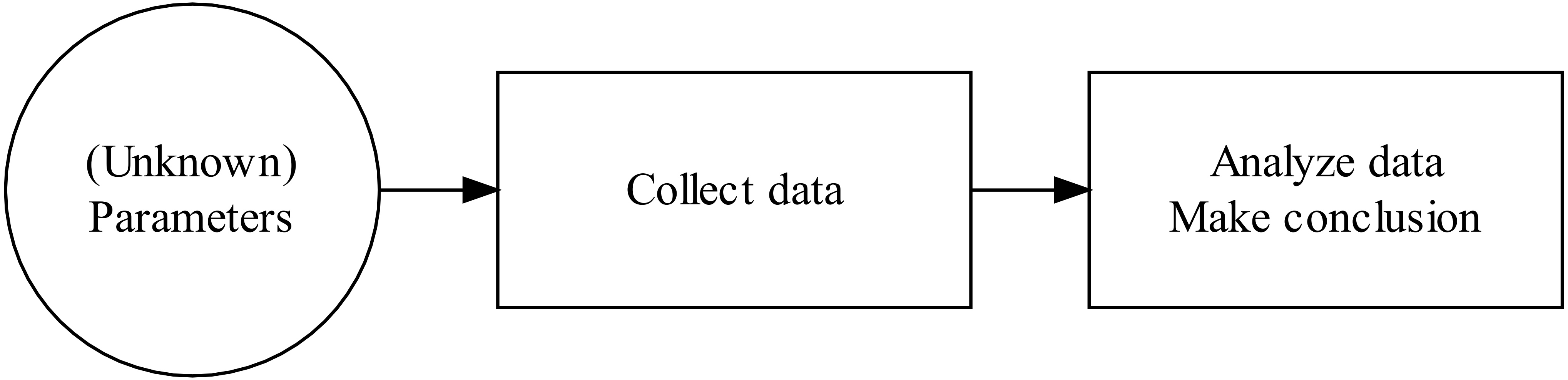
→ Examples of computing **optimism-corrected measures of performance** will be demonstrated in the breakout session.

Optimism is related to other common concepts in data analysis, such as:

- **over-fitting** or
- **bias-variance tradeoff** or
- **self-deception trap**

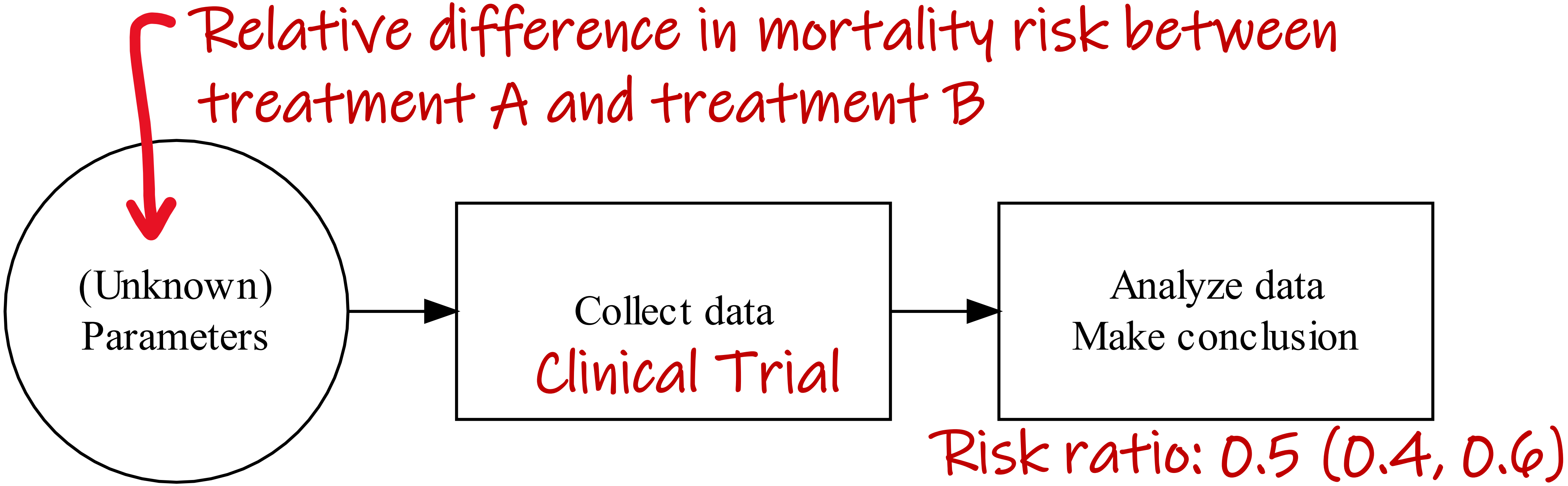
Course goal 2: Identify the operating characteristics of primary importance for prediction and inference

Inference

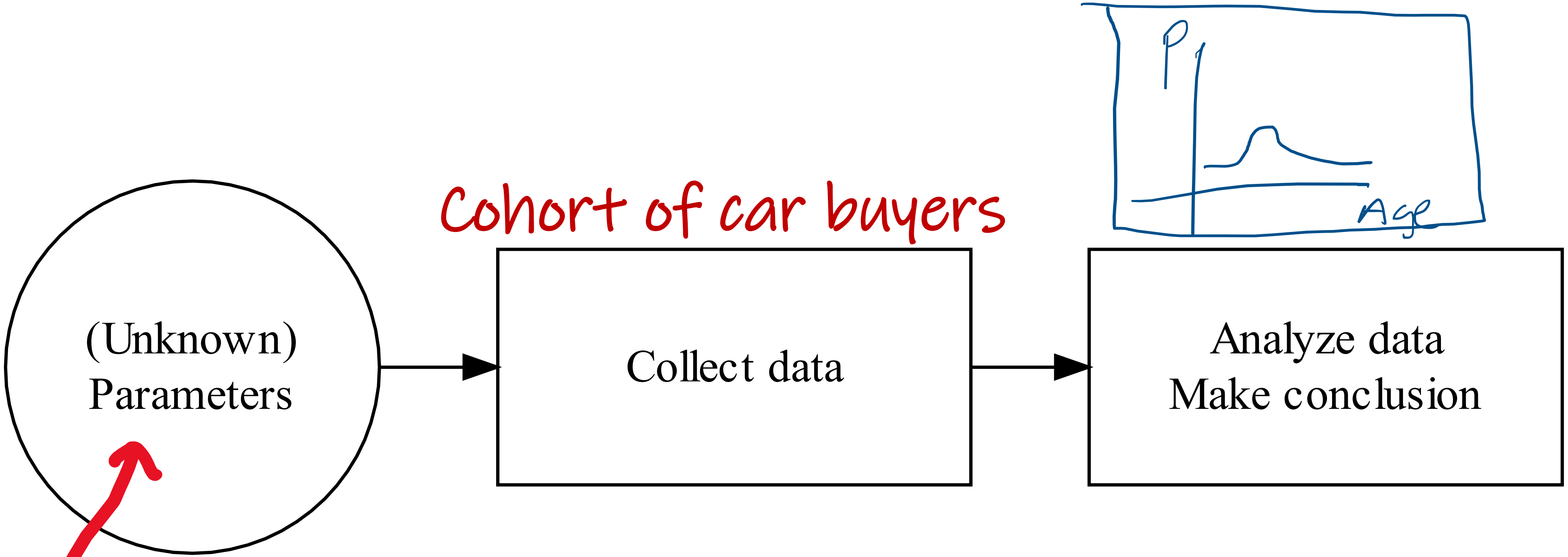


(Unknown)
Parameters
Collect data
Analyze data
Make conclusion

Inference


Relative difference in mortality risk between treatment A and treatment B
(Unknown) Parameters
Collect data
Clinical Trial
Analyze data
Make conclusion
Risk ratio: 0.5 (0.4, 0.6)

Inference



Association of age and propensity to purchase a Toyota Corolla

Key properties of a procedure (like collecting and analyzing data) are often called **operating characteristics**. Generally, one wants to know the **distribution** of an operating characteristic over repeated executions.

Key properties of a procedure (like collecting and analyzing data) are often called **operating characteristics**. Generally, one wants to know the **distribution** of an operating characteristic over repeated executions.

Operating characteristics are the **currency** by which we evaluate and compare data science procedures.

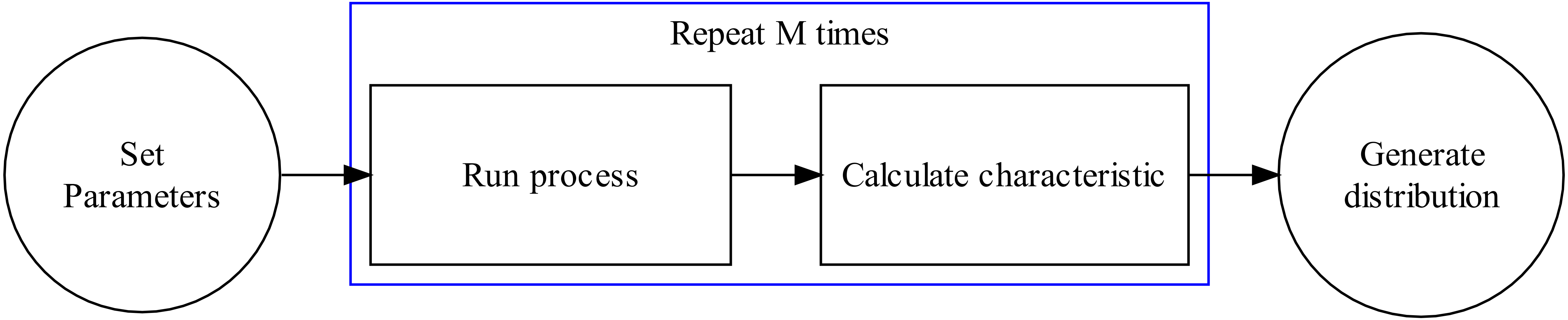
Example

- A data scientist claims to have developed a tool to identify college freshman that are highly likely to join the armed forces. **What operating characteristics would you like to know about the tool?**

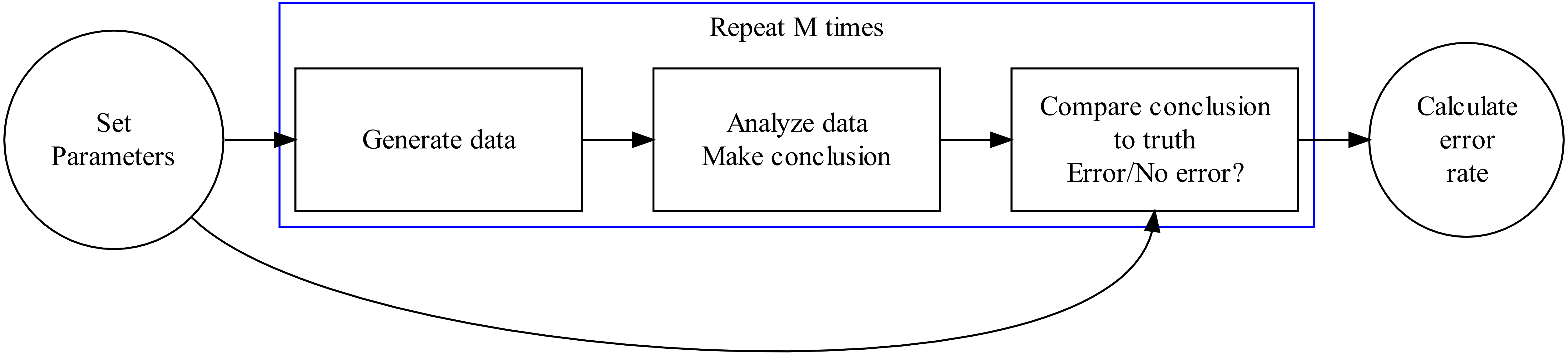
Example

- A data scientist develops an algorithm for estimating the probability that a credit card transaction is fraudulent or not. **What operating characteristics are important?**

Operating characteristics are premised on the classic "long-run" interpretation of probabilistic events. As such, they can be **simulated** by simply repeating the planned procedure and observing how often some event happens.

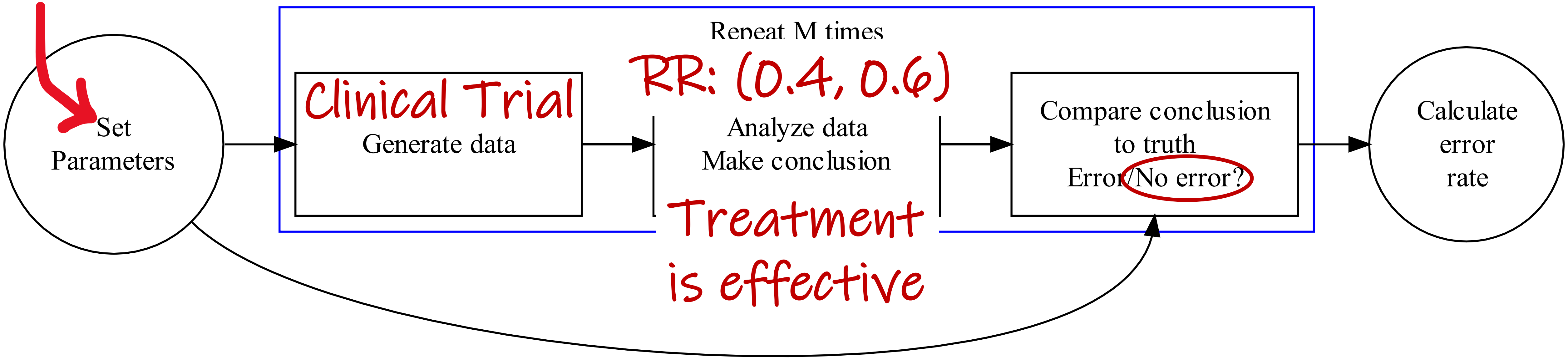


Example: Simulating operating characteristics for decision making

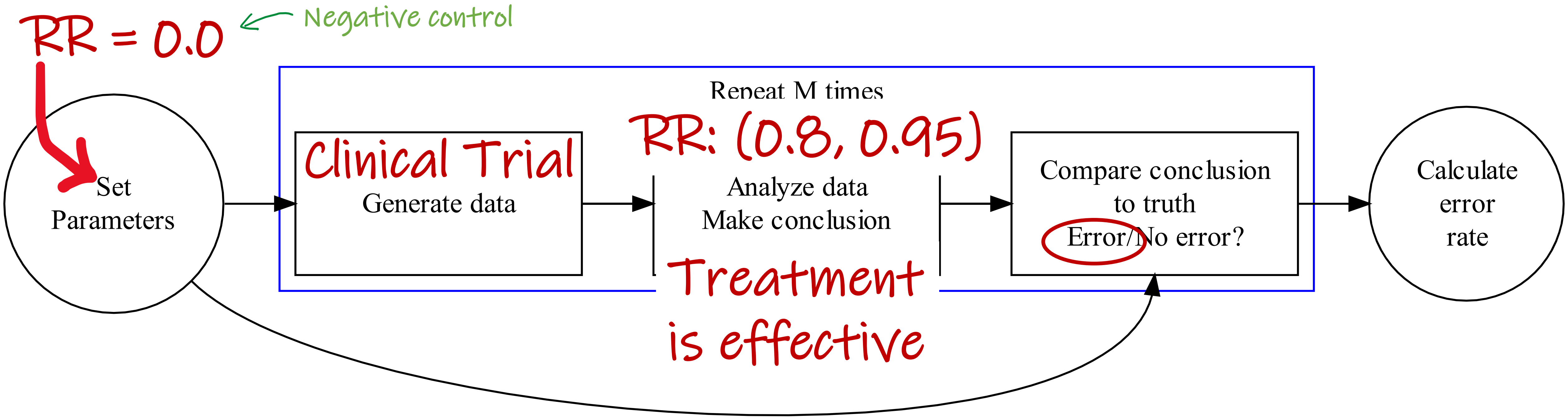


Relative risk of mortality comparing treatment A to placebo

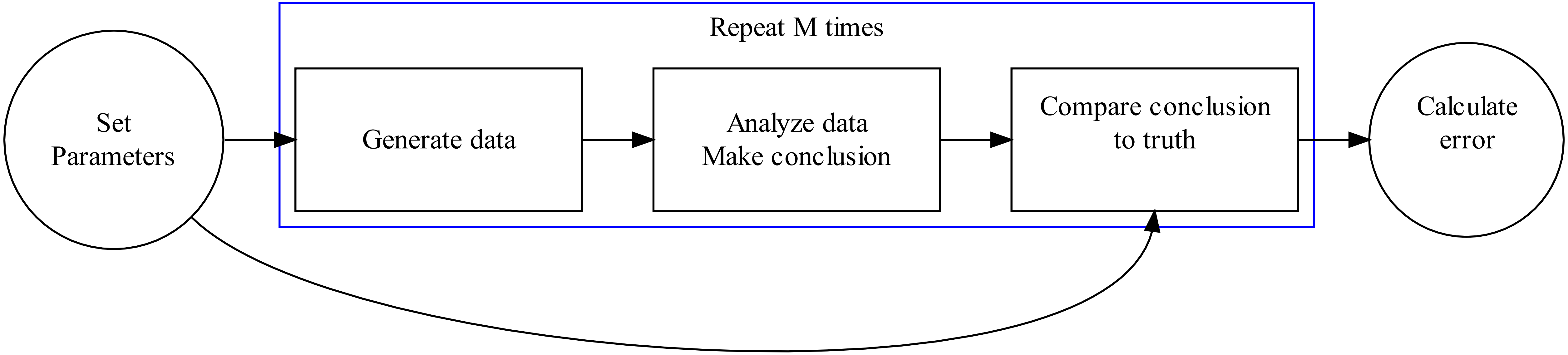
RR = 0.6

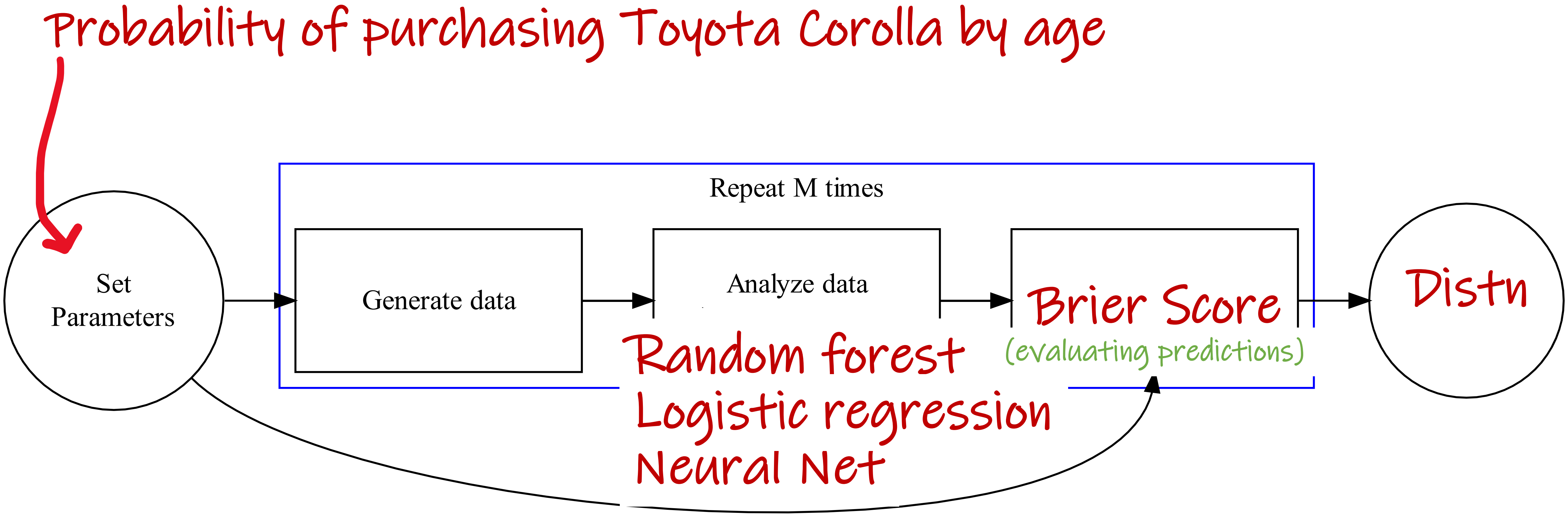


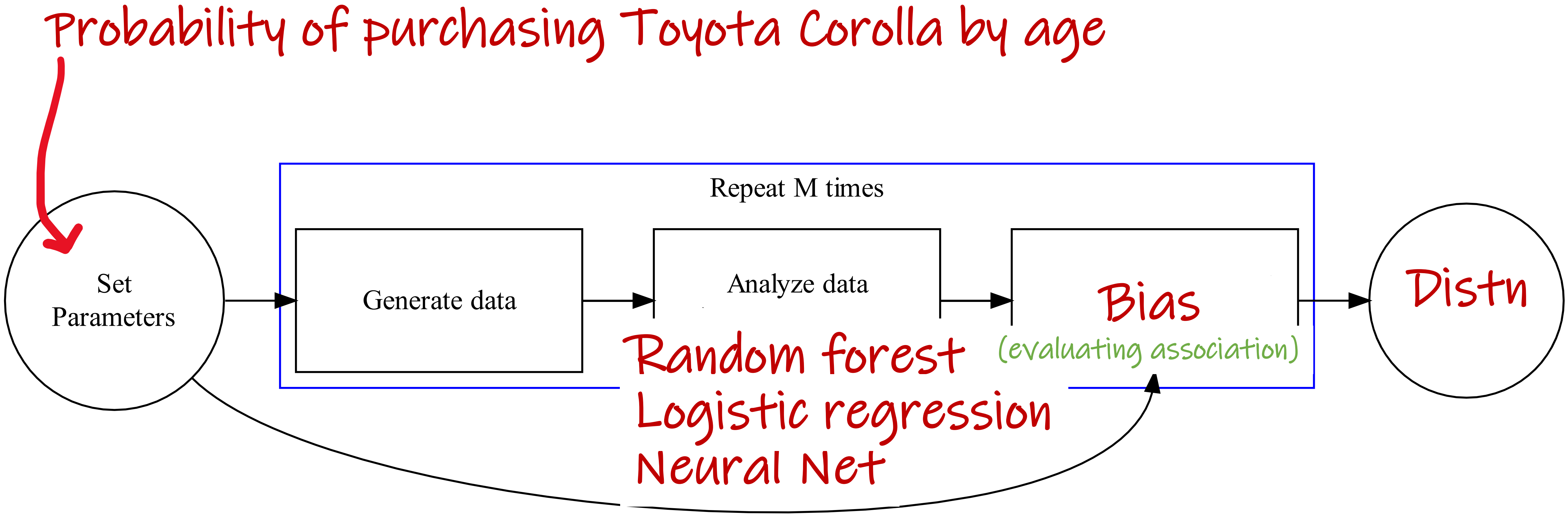
Relative risk of mortality comparing treatment A to placebo



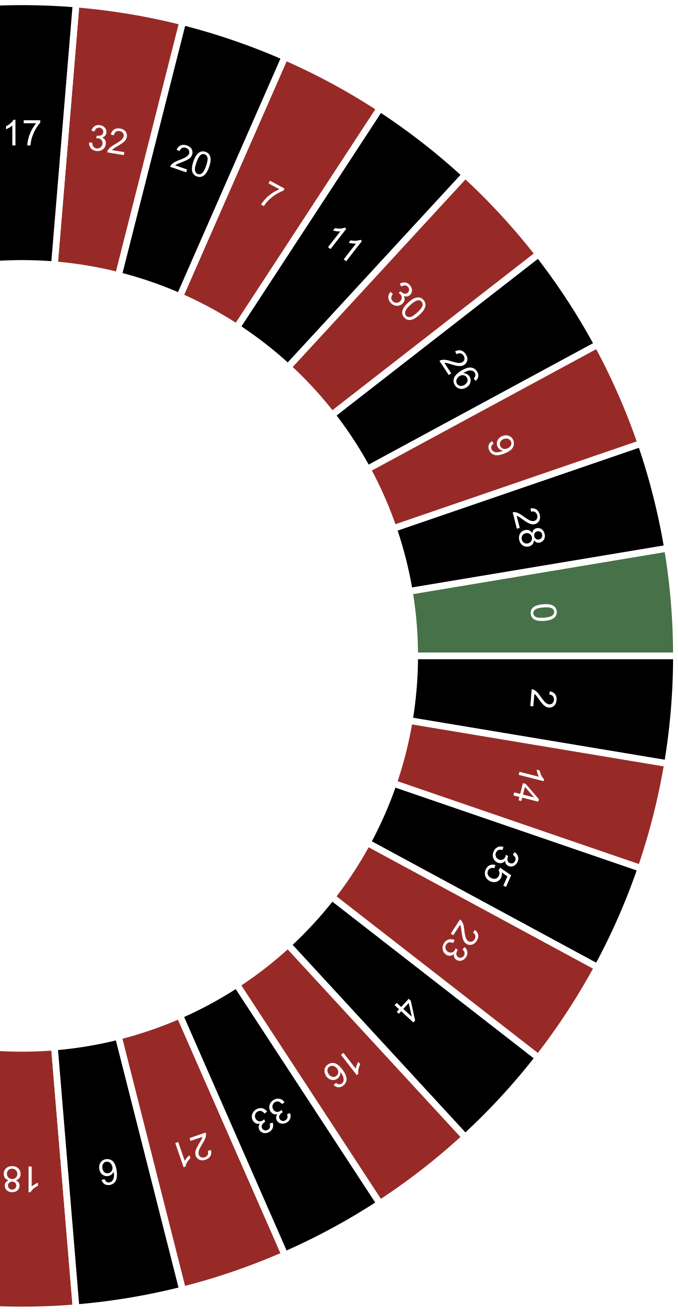
Example: Simulating operating characteristics for inference or prediction

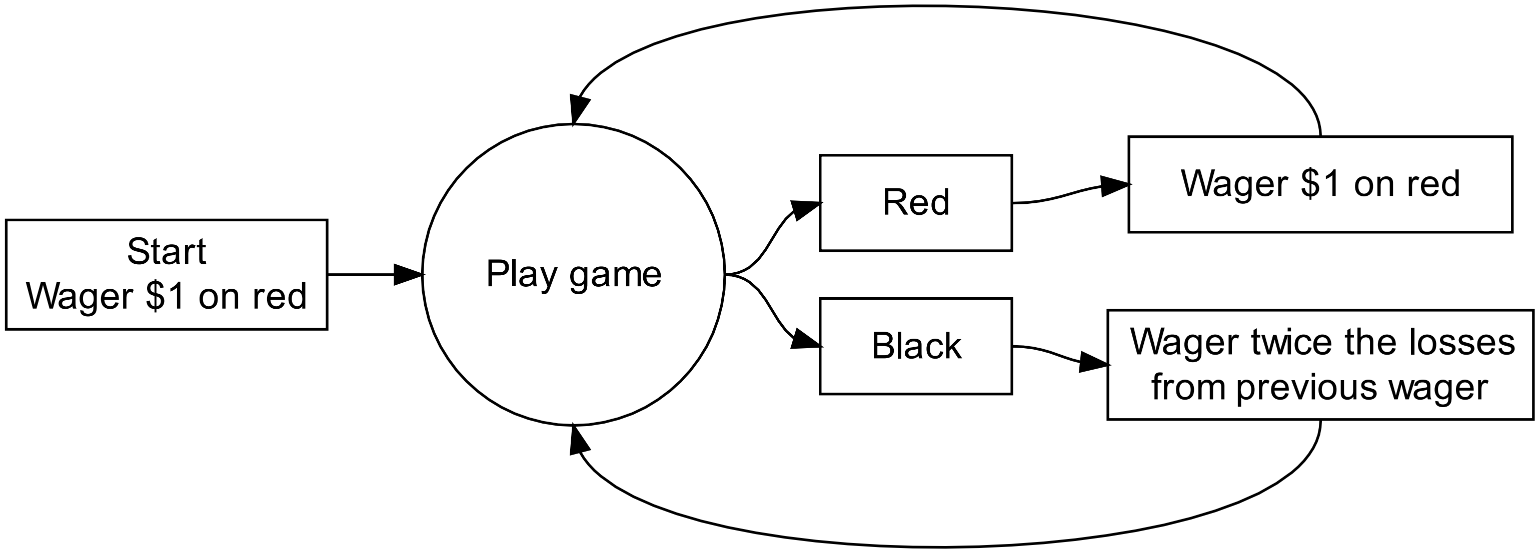


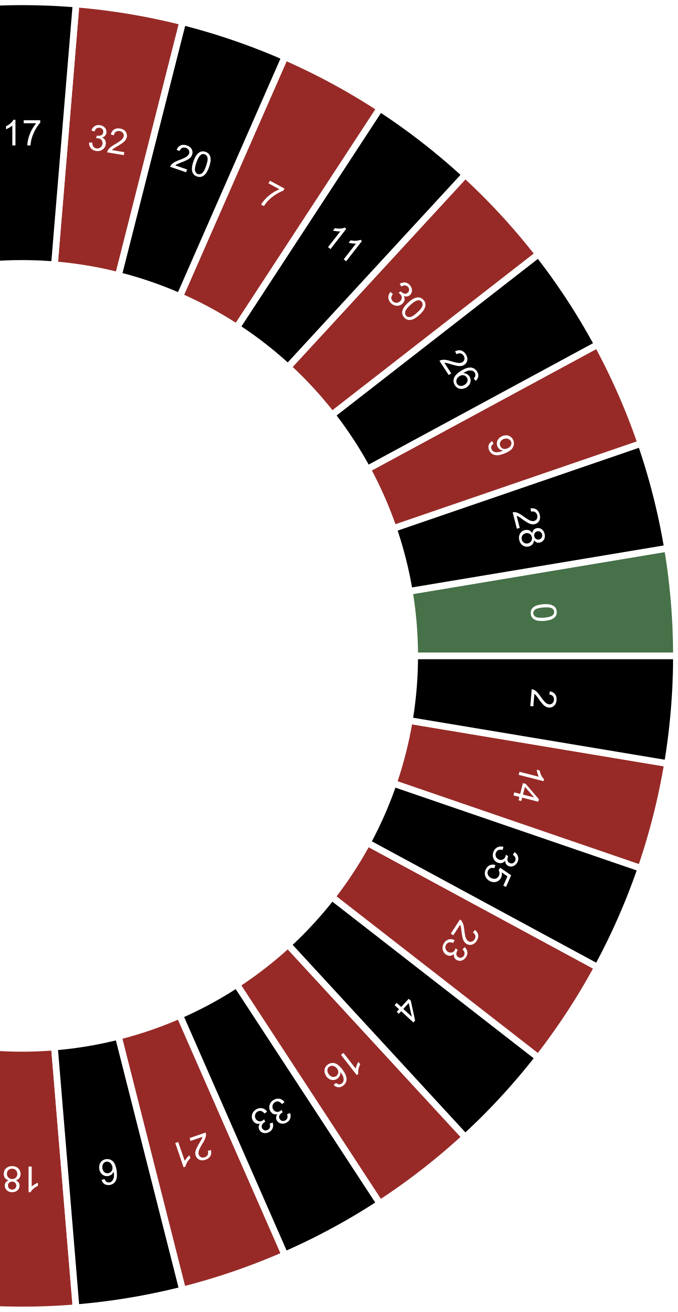

Probability of purchasing Toyota Corolla by age
Set Parameters
Repeat M times
Generate data
Analyze data
Brier Score
(evaluating predictions)
Random forest
Logistic regression
Neural Net
Distn


Probability of purchasing Toyota Corolla by age
Set Parameters
Repeat M times
Generate data
Analyze data
Bias
(evaluating association)
Distn
Random forest
Logistic regression
Neural Net

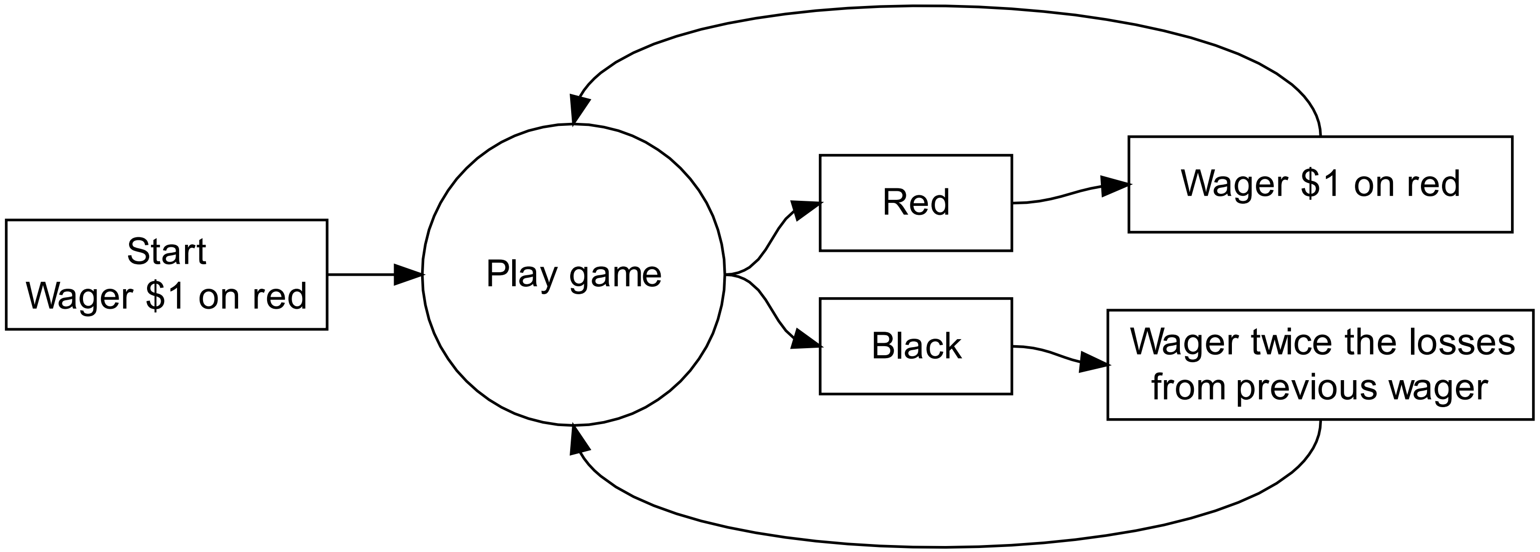
Roulette Example



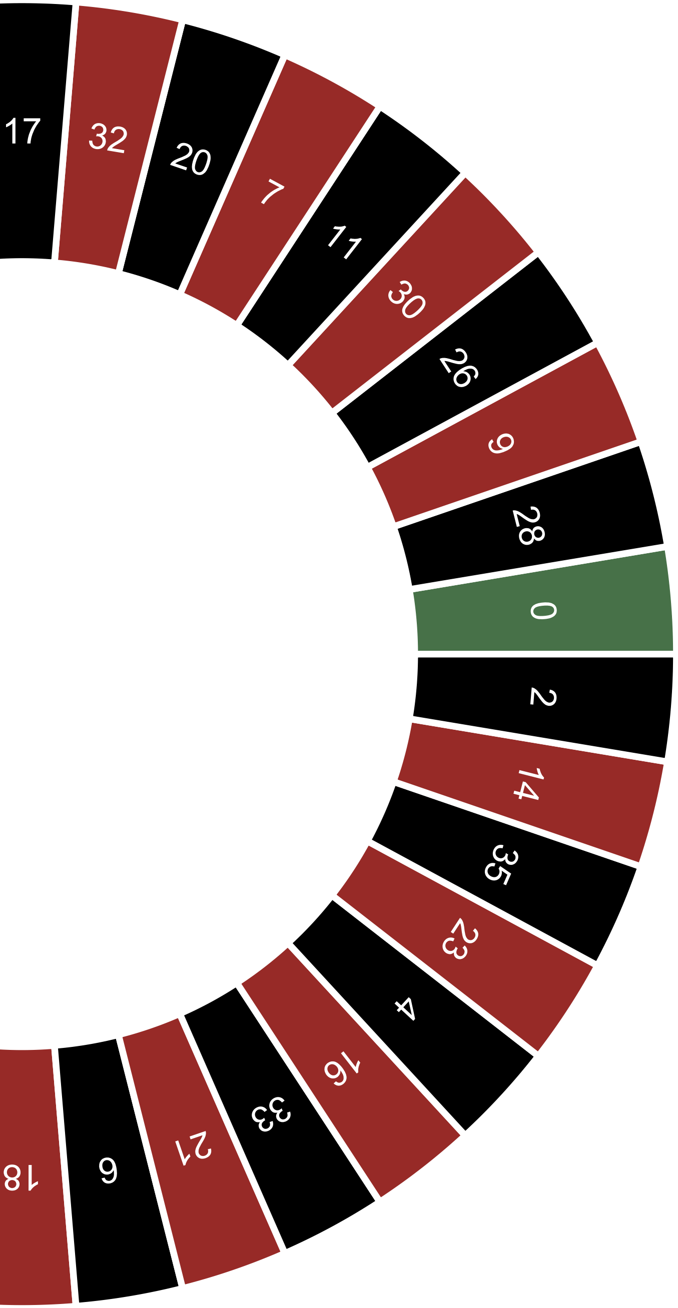




Roulette Example



Does this betting strategy work?



Roulette Example

Setup, code, and video solution at:
tgstewart.xyz/**roulette**

Course goal 2: Identify the operating characteristics of primary importance for prediction and inference

Prediction	Inference
Discrimination	Bias
Calibration	Coverage

Stability

Course goal 4: Recognize the pitfalls of variable selection techniques when constructing models for inference

Challenge

Standard errors (and consequently confidence intervals) generated from a regression model or machine learning algorithm usually assume the predictor variables were selected ***a priori***, without reference to the data.

Challenge

Standard errors (and consequently confidence intervals) generated from a regression model or machine learning algorithm usually assume the predictor variables were selected ***a priori***, without reference to the data.

Data driven variable selection prior to model fitting and inference introduces **additional variability** that is not captured with standard methods of computing **standard errors**.

Challenge

Standard errors (and consequently confidence intervals) generated from a regression model or machine learning algorithm usually assume the predictor variables were selected ***a priori***, without reference to the data.

Data driven variable selection prior to model fitting and inference introduces **additional variability** that is not captured with standard methods of computing **standard errors**.

Data driven variable selection may also introduce **bias** to **parameter estimates**.

Challenge

Standard errors (and consequently confidence intervals) generated from a regression model or machine learning algorithm usually assume the predictor variables were selected *a priori*, without reference to the data.

Mostly referring to stepwise procedures.

Data driven variable selection prior to model fitting and inference introduces additional variability that is not captured with standard methods of computing standard errors.

Data driven variable selection may also introduce bias to parameter estimates.

Solutions

See Dr. Jeffrey Blume’s slides.

¯_(ツ)_/¯

Course goal 3: Simulate operating characteristics for simple prediction and inference models https://intobioinformatics.wordpress.com/2018/12/25/optimism-corrected-bootstrapping-a-problematic-method/

Optimism corrected bootstrapping: a problematic method

Search …

December 25, 2018

There are lots of ways to assess how predictive a model is while correcting for overfitting. In Caret the main methods I use are leave one out cross validation, for when we have relatively few samples, and k fold cross validation when we have more. There also is another method called 'optimism corrected bootstrapping', that attempts to save statistical power, by first getting the overfitted result, then trying to correct this result by bootstrapping the data to estimate the degree of optimism. A few simple tests in Caret can demonstrate this method is bunk.

This is a very straightforward method, just add random variables from a normal distribution to the ground truth iris labels. We should find our AUC (area under ROC curve) is about 0.5. Yet for optimism corrected bootstrap it gives a positive result regardless of whether the predictors are just noise or not. Let's just run that test:

This is called a sensitivity analysis for the uninitiated, I am just increasing number of random noise features (z) and binding them to the real labels in an iterative manner.

```
library(caret)
allresults <- matrix(ncol=2,nrow=200)
i = 0
for (z in seq(10,2000,10)){

  i = i + 1

  # select only two species
```

Why this example

- If you had asked me how to estimate **out-of-sample performance** for a logistic regression, I would have told you **[as I've done again today]** to consider **optimism corrected measures** of model performance.

Why this example

- If you had asked me how to estimate **out-of-sample performance** for a logistic regression, I would have told you **[as I've done again today]** to consider **optimism corrected measures** of model performance.
- In late December 2018, I was made aware of an interesting, online discussion about the limits of optimism corrected measures when the number of predictors is large.

The Problem

- **Optimism corrected AUC** did not seem to work with a large number of predictors in the **negative control setting,** where all of the predictors were just noise.

https://intobioinformatics.wordpress.com/2018/12/25/optimism-corrected-bootstrapping-a-problematic-method/

Optimism corrected bootstrapping: a problematic method

There are lots of wa
ting. In Caret the ma
have relatively few s
also is another meth
save statistical pow
result by bootstrapp

Part 2: Optimism corrected bootstrapping is definitely bias, further evidence

Some peo
strapping'
the numbe
blog post)
the intere

Part 3: Two more implementations of optimism corrected bootstrapping show clear positive results bias

Previousl
bootstrap
the code

This time
logistic re

Part 4: Why does bias occur in optimism corrected bootstrapping?

In the previous parts of the series we demonstrated a
mism corrected bootstrapping by simply adding rand
problem is due to an 'information leak' in the algorith
test datasets are not kept seperate when estimating
optimism, under some conditions, can be very under
code, it is pretty straightforward to understand then
originates.

Part 5: Code corrections to optimism corrected bootstrapping series

The truth is out there
The previous post ex
bootstrapping (a met
with increasing p (co
lications. However, th
ous post, 1 has a sligl

Part 6: How not to validate your model with optimism corrected bootstrapping

When evaluating a machine learning model if the same data is used to train and test the model this results in overfitting. So the model performs much better in predictive ability than it would if it was applied on completely new data, this is because the model uses random noise within the data to learn from and make predictions. However, new data will have different noise and so it is hard for the overfitted model to predict accurately just from noise on data it has not seen.

The blog post spurred several statisticians to simulate the operating characteristics in order to understand the method’s limits

http://hbiostat.org/doc/simval.html

Comparison of Strategies for Validating Binary Logistic Regression Models

Frank Harrell

2018-12-29

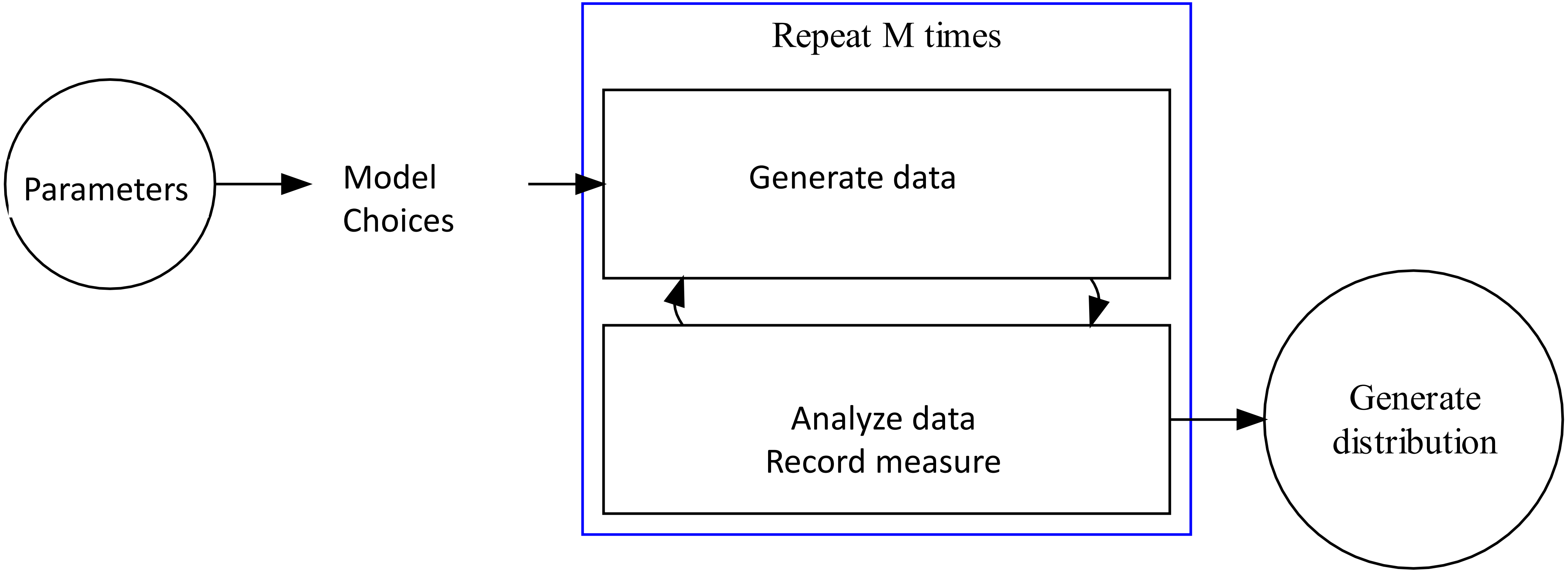
Simulation Method

For each of 400 simulations generate a training sample of 500 observations with p predictors (p=15, 30, 60, 90) and a binary reponse. The predictors are independently U(-0.5,0.5). The response is sampled so as to follow a logistic model where the intercept is zero and the regression coefficients have each of two patterns. First, all coefficients are set to 0.0 so that the true predictive model has no predictive discrimination ability ($D_{xy} = 0, c =$ AUROC $= 0.5$). Then regression coefficients were uniformly spaced between -1 and 1, multiplied by a scaling factor that is < 1 when the number of predictors p is 30 and > 1 when p > 30. The “gold standard” is the predictive ability of the fitted model on a test sample containing 50,000 observations generated from the same population model. The task of a validation method is to recover this gold standard.

What was learned?

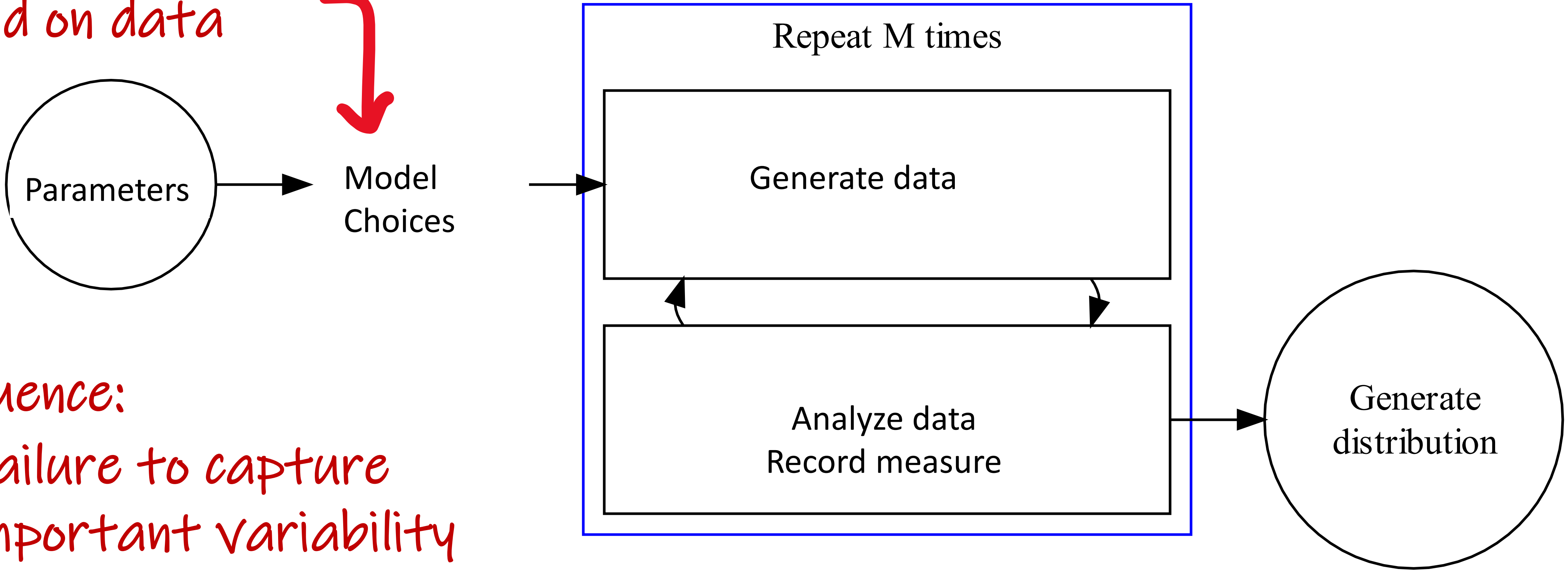
- The performance of bootstrap optimism correction depended on the measure of model performance
    - For AUC related measures, the ratio of predictors to sample size is an important factor
        - If N < 5*P, another measure of out-of-sample AUC is needed.

Big picture


Parameters
Model
Choices
Repeat M times
Generate data
Analyze data
Record measure
Generate
distribution

Big picture

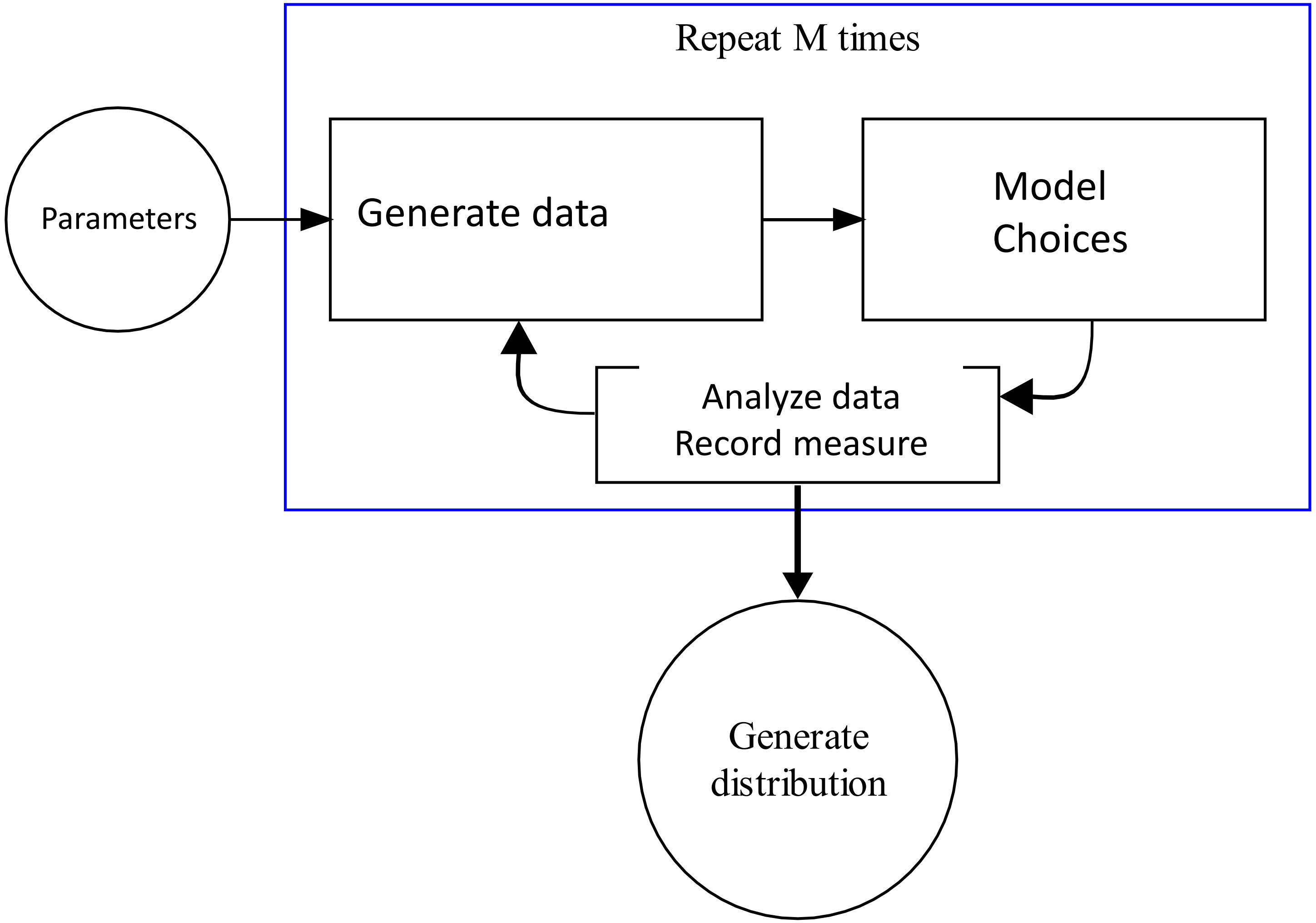
Pitfall: Model choices depend on data



Consequence:

Failure to capture important variability

Capture variability



Examples
+ Number of clusters
+ Variables in model
+ Form of predictors

Simulation is a great approach to estimating trial designs

Trial design characteristics:

+ Power

+ Assertion rates

+ Expected Sample Size

+ Futility