Data Divers

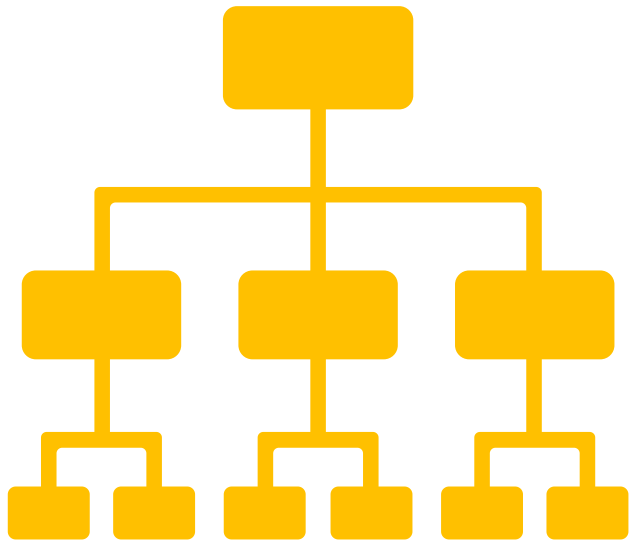
Data Science – Yellow Belt

Lecture 2

by Sebastian Sauer

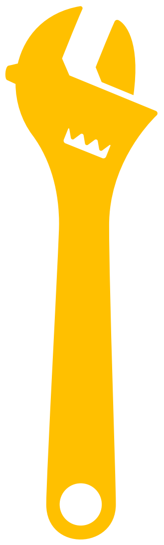
Statistical Learning: What's that?

Big Picture



Lecture 1

Lecture 2



Lecture 3

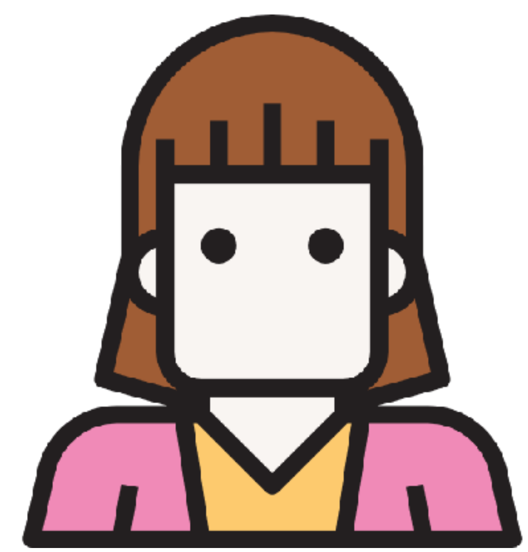
Learning goals

Let's explore the mathematical intricacies!

Wolfi

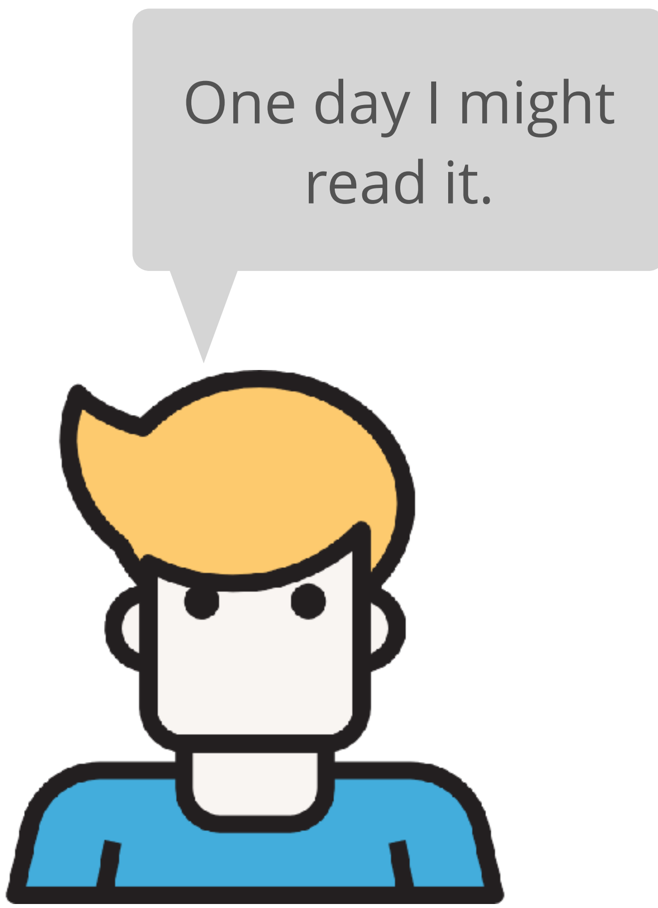
OK, but I'll throw in some background where and when needed.

No, this time we'll get our hands dirty with state-of-the-art case-studies.

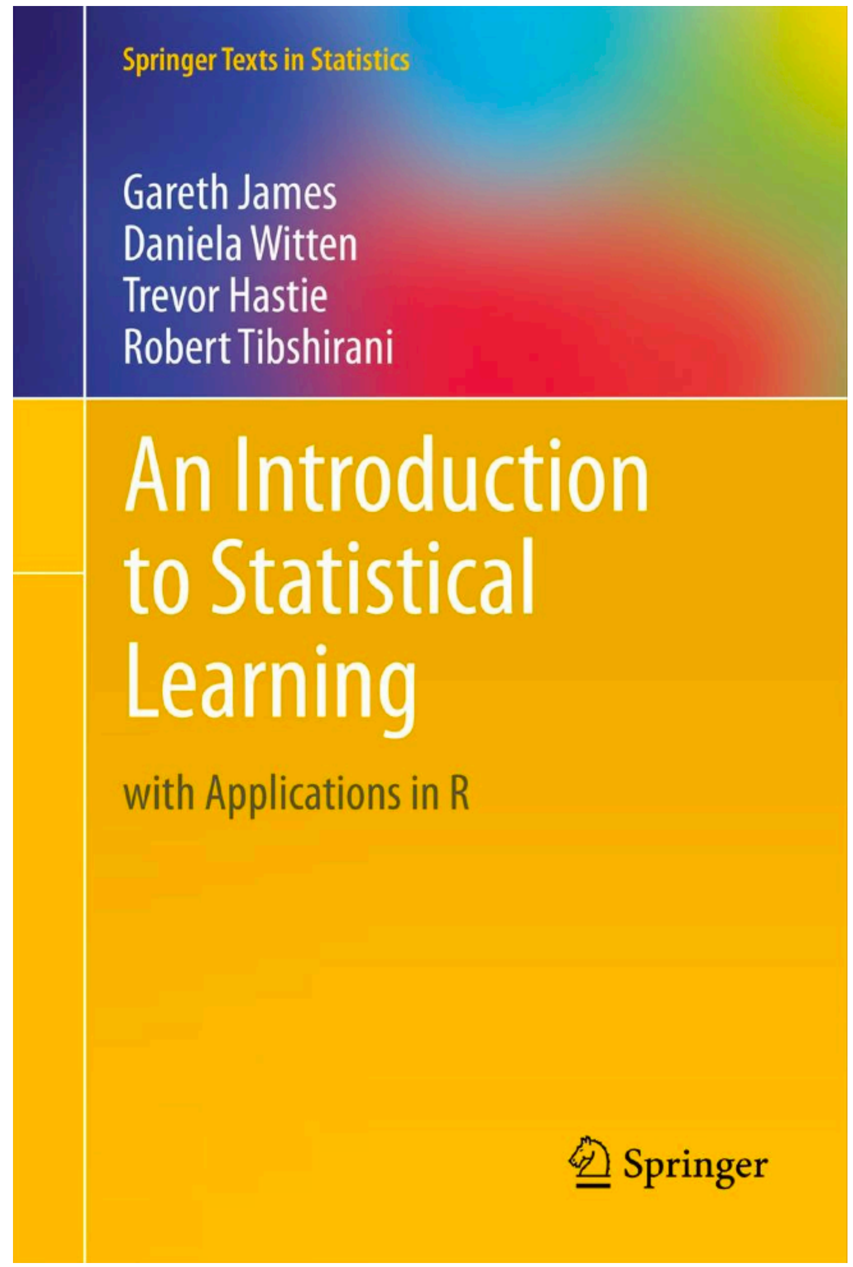


Angi

The standard source of knowledge



Don



[ebook freely available](https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf)

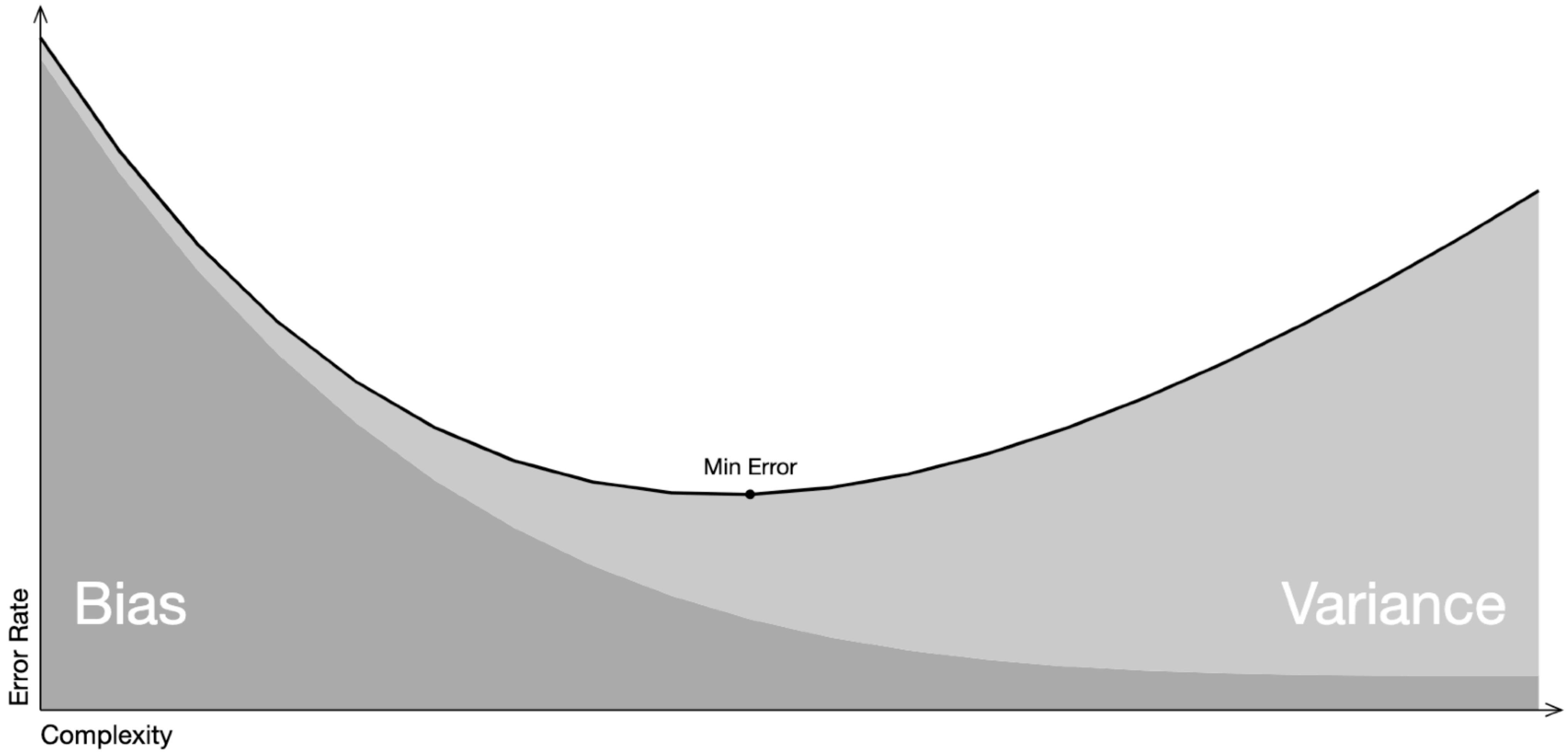
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

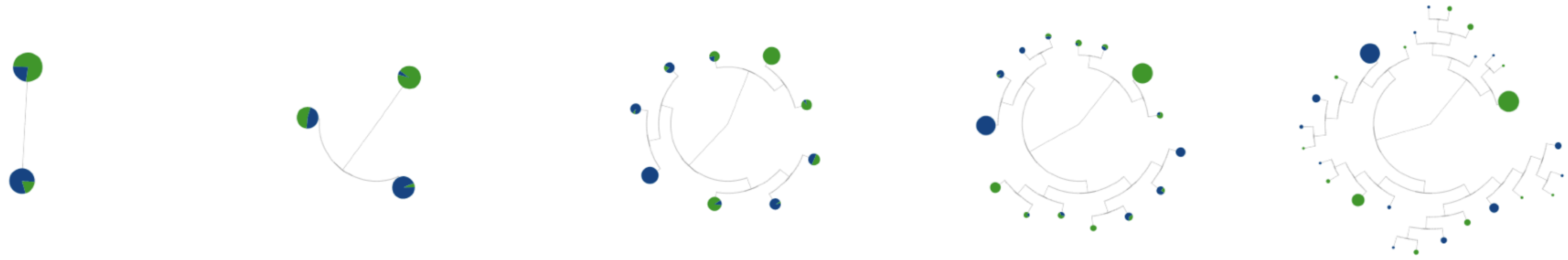
Building a machine learning pipeline, 101

- In break-out groups, work your way through this demonstration of a machine learning pipeline.
- Answer the following questions:
  1. What are model parameters?
  2. Which models tend to exhibit a strong bias?
  3. Why do overly-complex exhibit high variance?
  4. What's a way to balance the bias-variance tradeoff?
- Feel free to double check part 1 of this demonstration.

http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

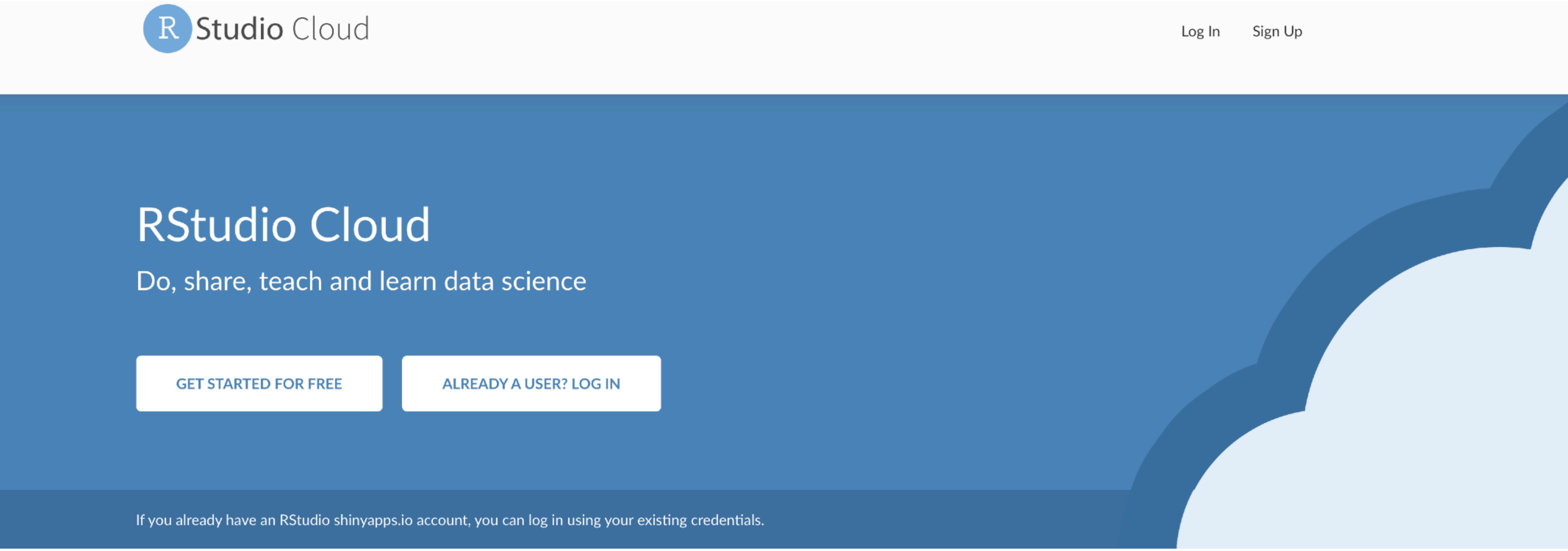
Explain bias-variance tradeoff in this diagram





http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

Open your RStudio Cloud account

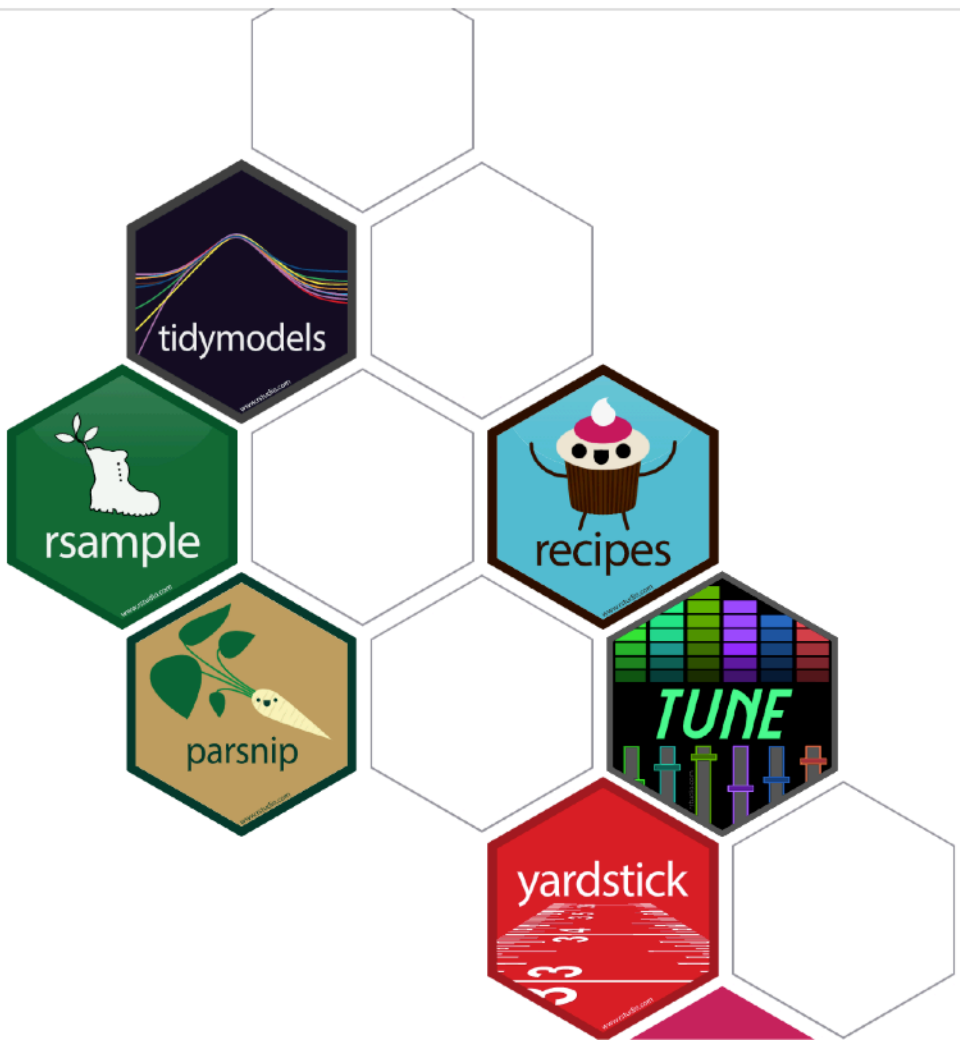


https://rstudio.cloud/

We'll use the „tidymodels" data science framework

Tidymodels

PACKAGES GET STARTED LEARN HELP CONTRIBUTE



TIDYMODELS

The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.

Install tidymodels with:

```
install.packages("tidymodels")
```

https://www.tidymodels.org/

Here's our tutorial/case study

GET STARTED

Stuck? Confused? Ask for help.

1 Build a model

TIDYMODELS PACKAGES: broom, parsnip

INTRODUCTION

How do you create a statistical model using tidymodels? In this article, we will walk you through the steps. We start with data for modeling, learn how to specify and train models with different engines using the parsnip package, and understand why these functions are designed this way.

To use code in this article, you will need to install the following packages: broom.mixed, readr, rstanarm, and tidymodels.

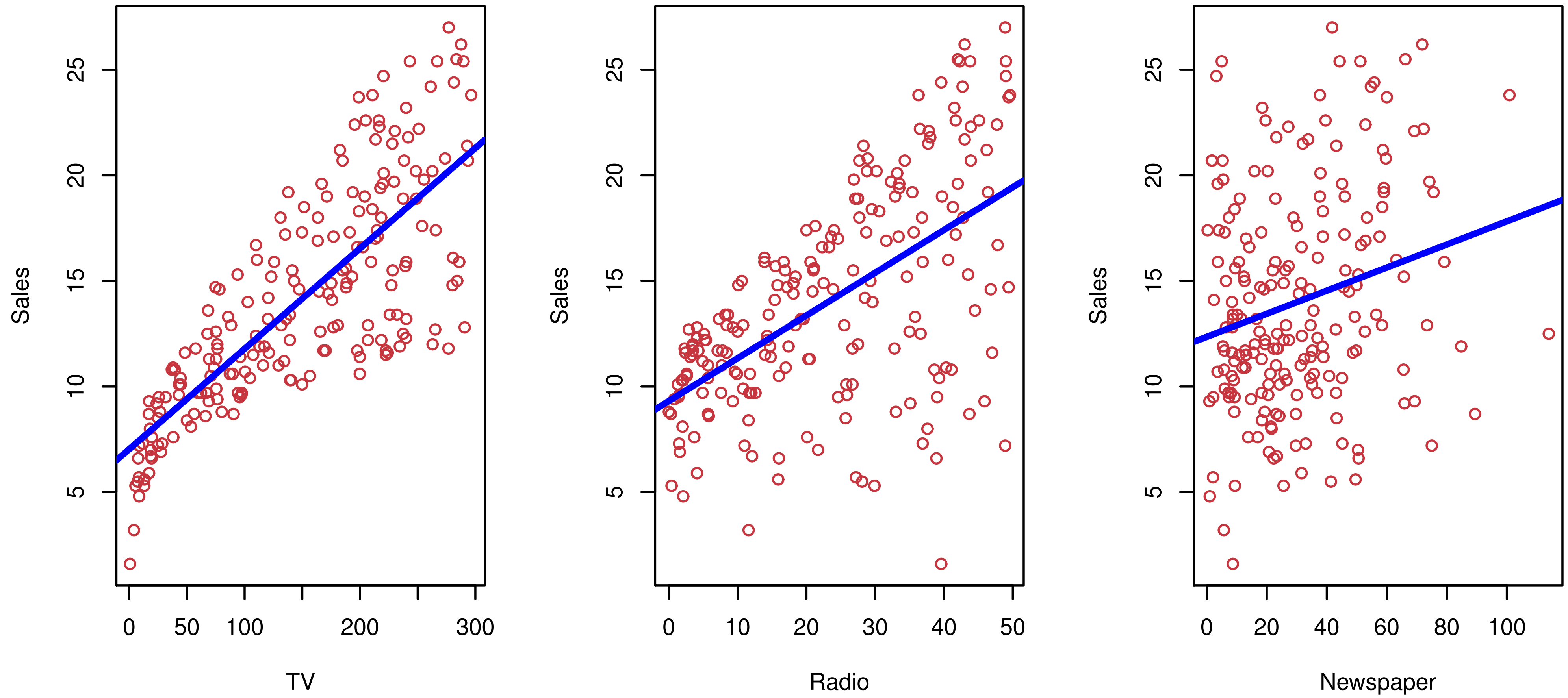
https://www.tidymodels.org/start/models/

Statistical Learning: Finding the pattern of Y and X

$$Y = f(X) + e$$

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Finding patterns is our game



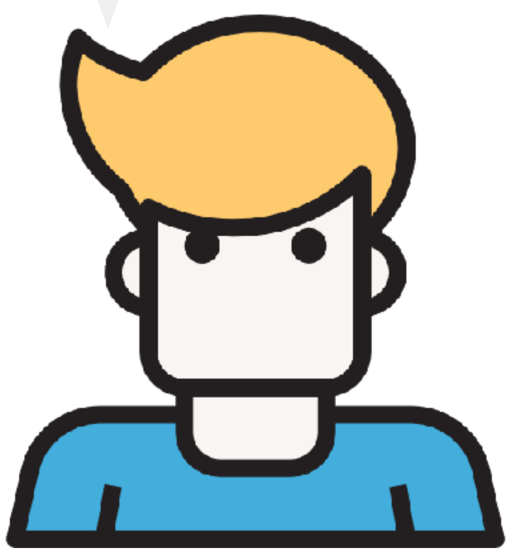
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Why estimating f ?

Prediction

Explanation

Who cares about „why“ as long you get accurate predictions!



Don

We need to understand what’s going on.

Wolfi

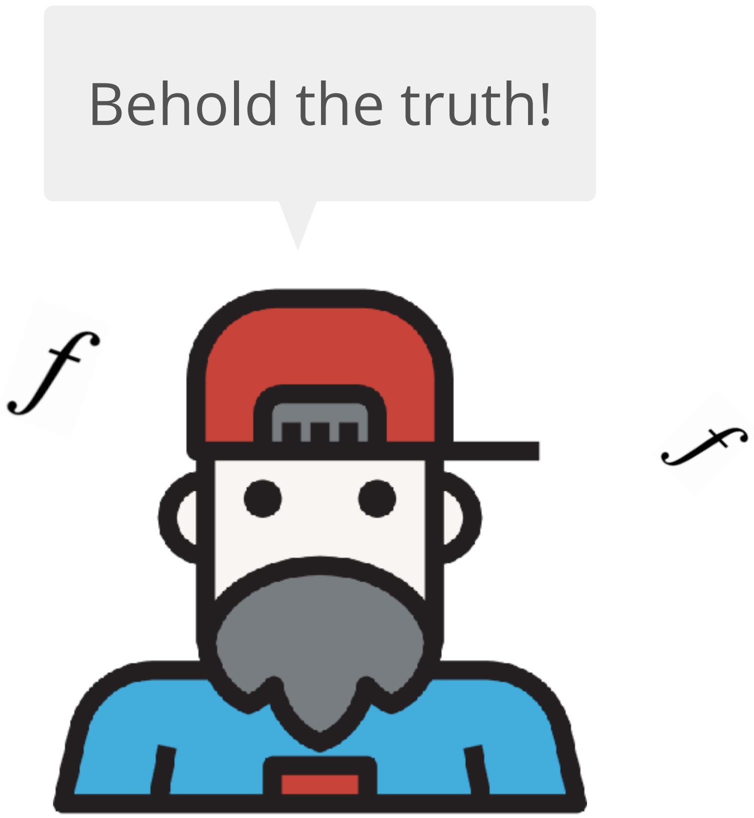
Two types of errors: reducible and non-reducible

$$e = e_r + e_{nr}$$

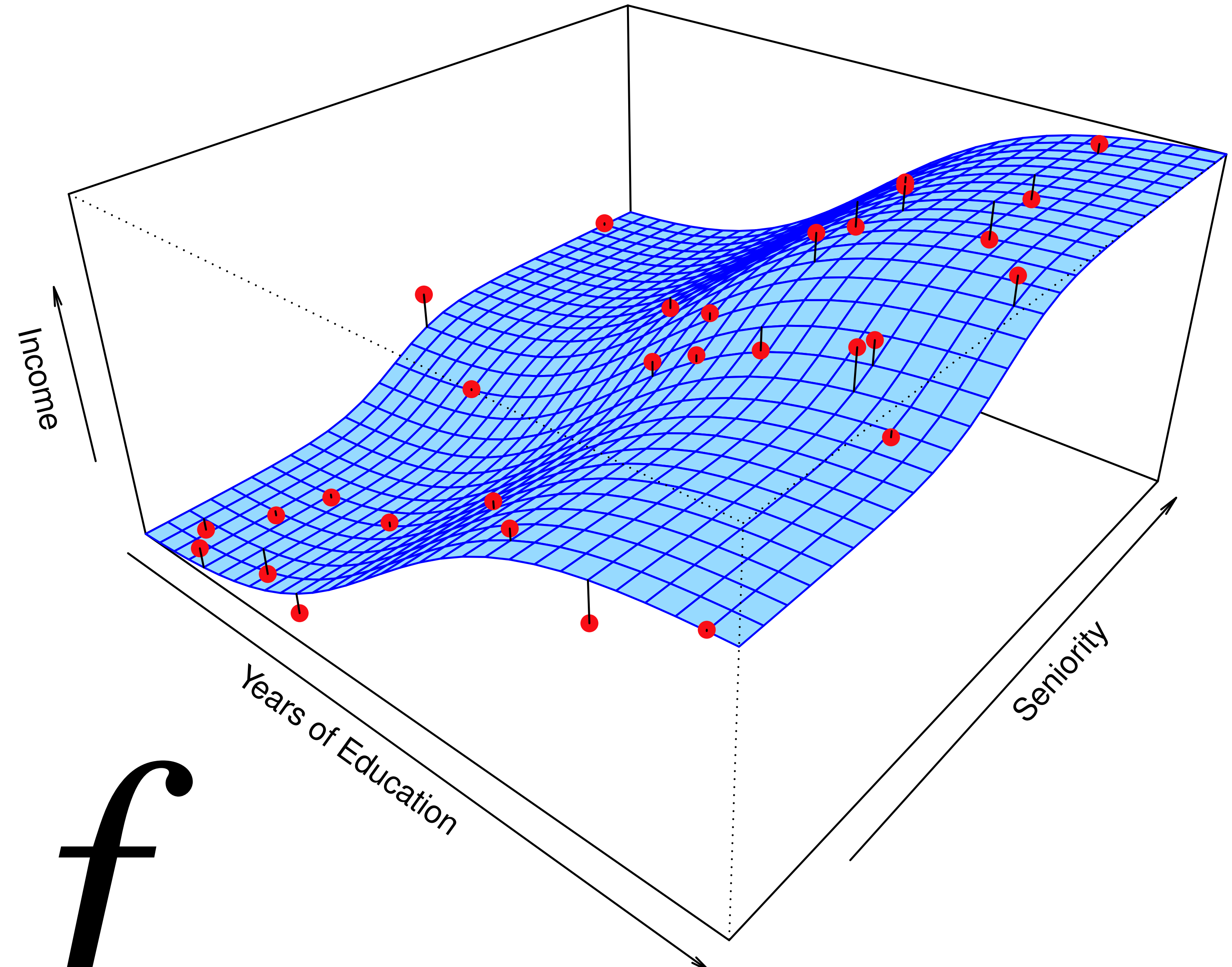
$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = \underbrace{E[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{Var(\epsilon)}_{\text{non-reducible}}$$

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Here's the non-reducible error



Wolfi

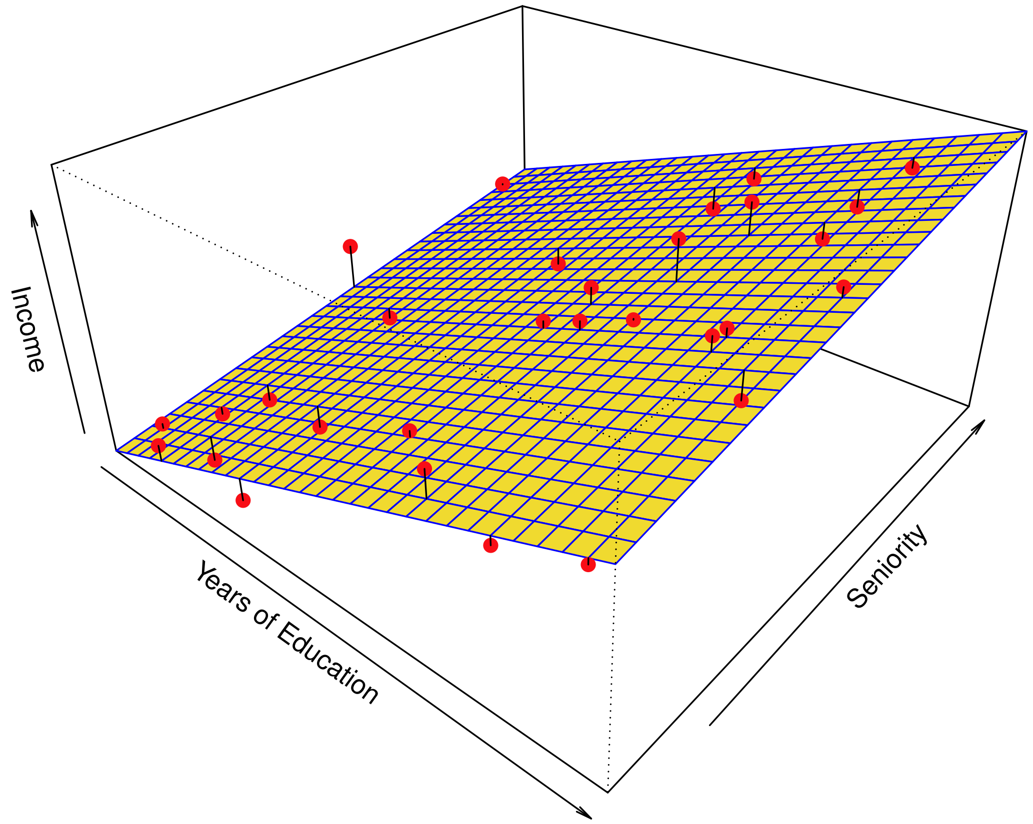


f

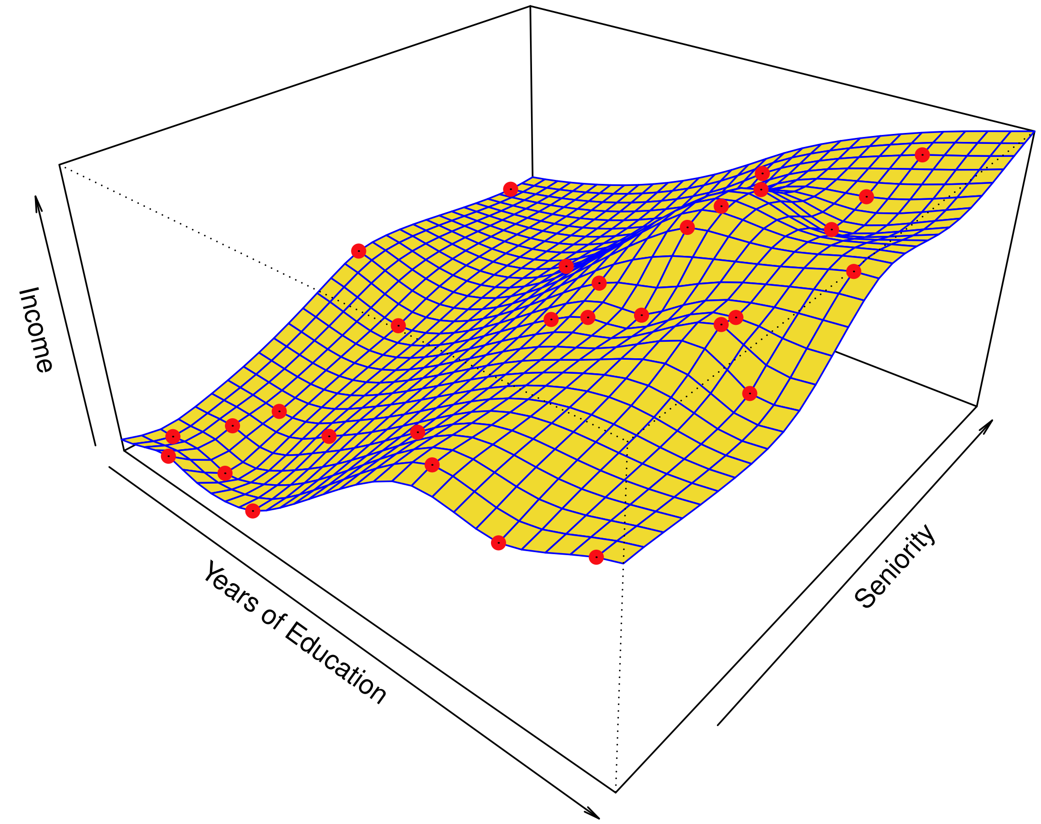
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Parametric and non-parametric models

Parametric model



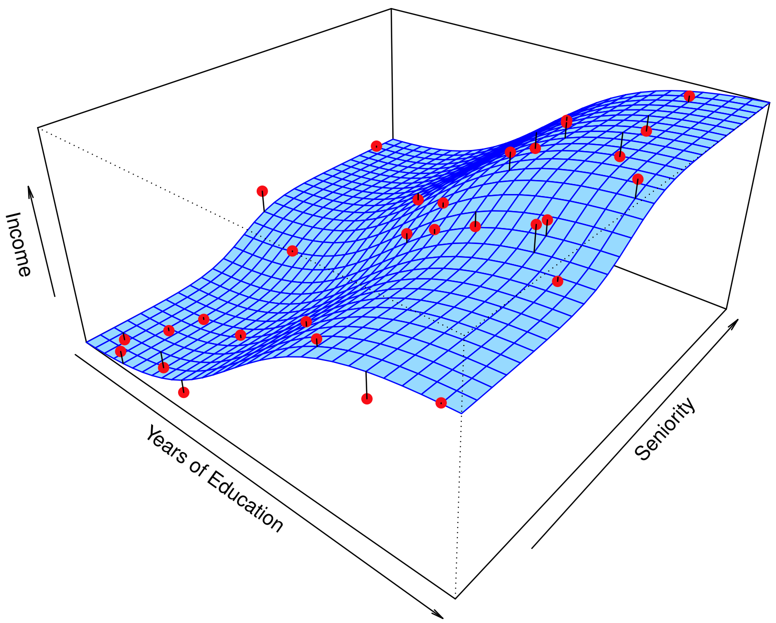
Non-parametric Model

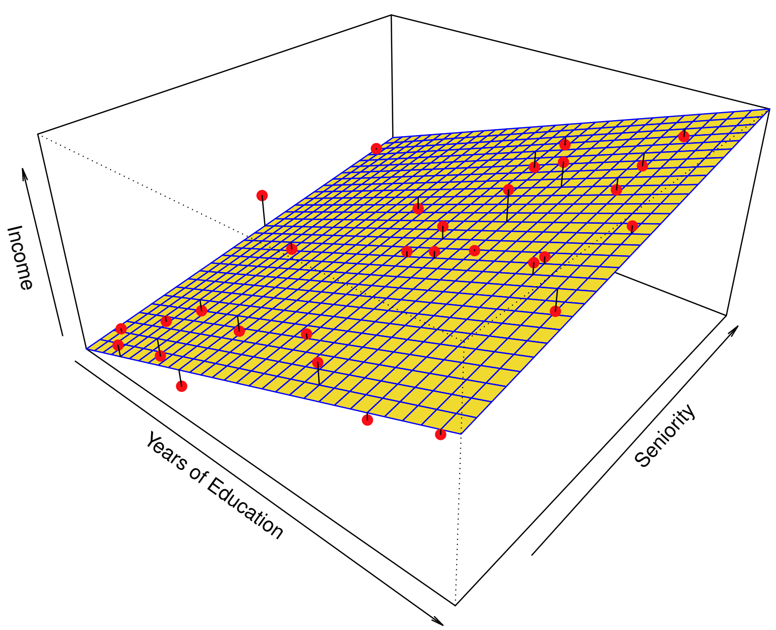


James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

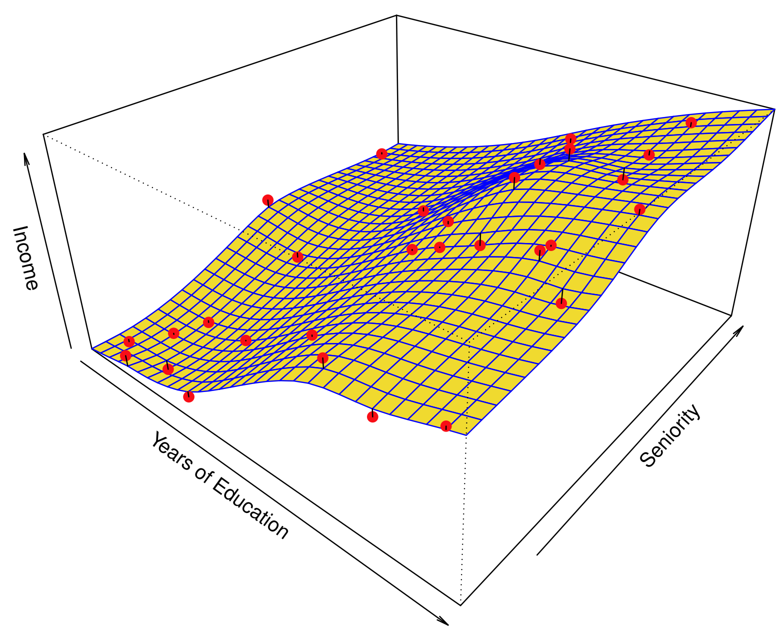
From too simple to too complex

$$f$$

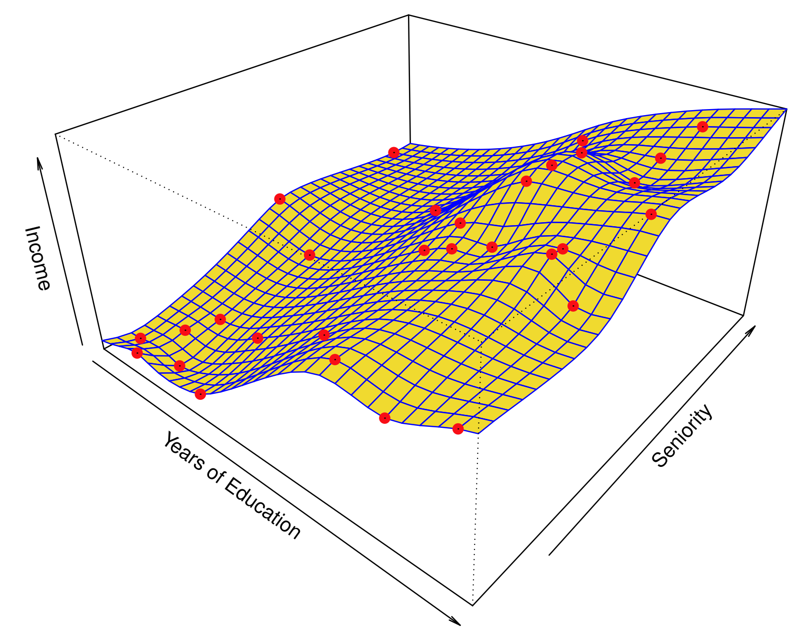




$$\hat{f}_1$$



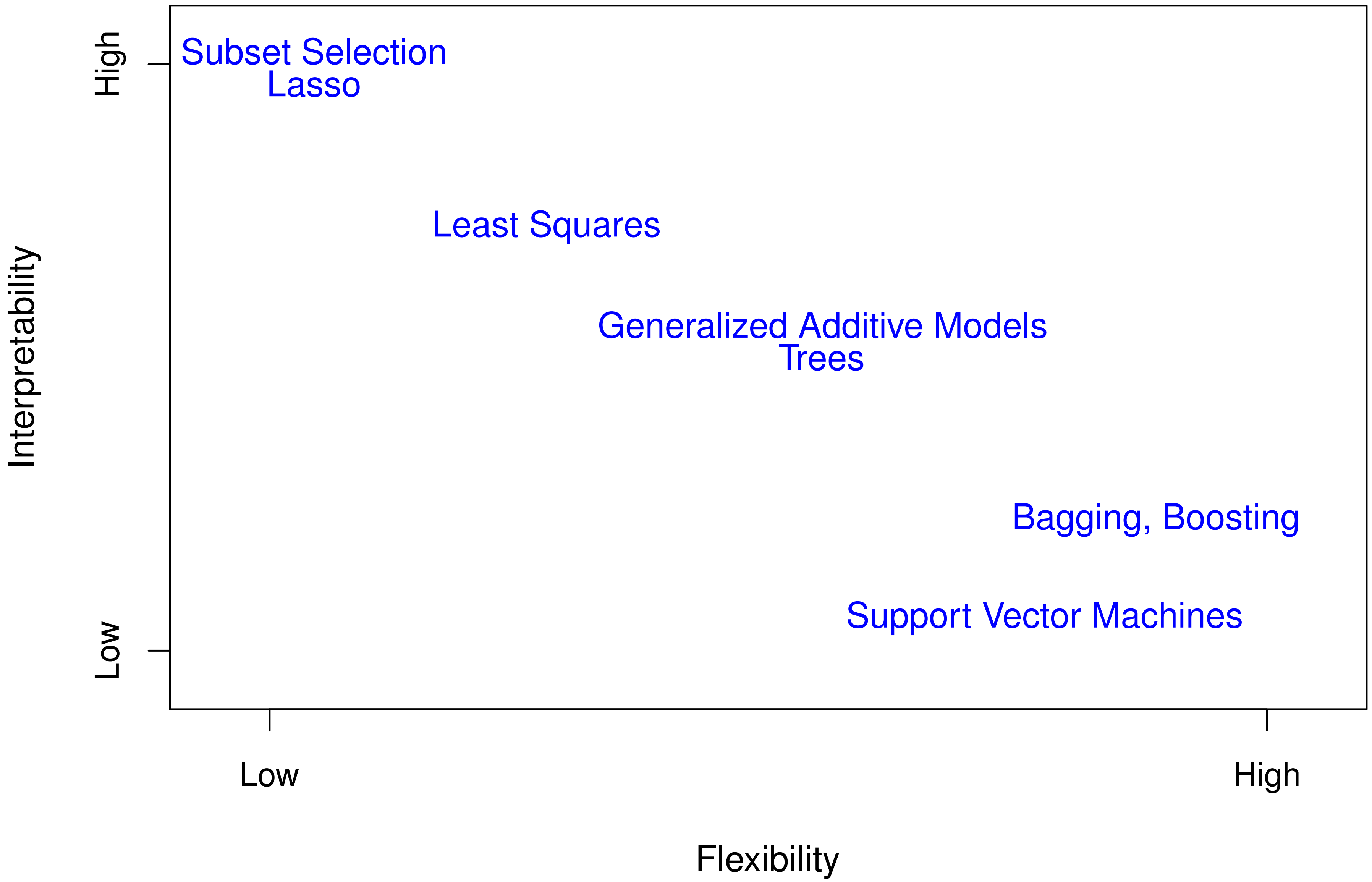
$$\hat{f}_2$$



$$\hat{f}_3$$

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Interpretability vs. flexibility

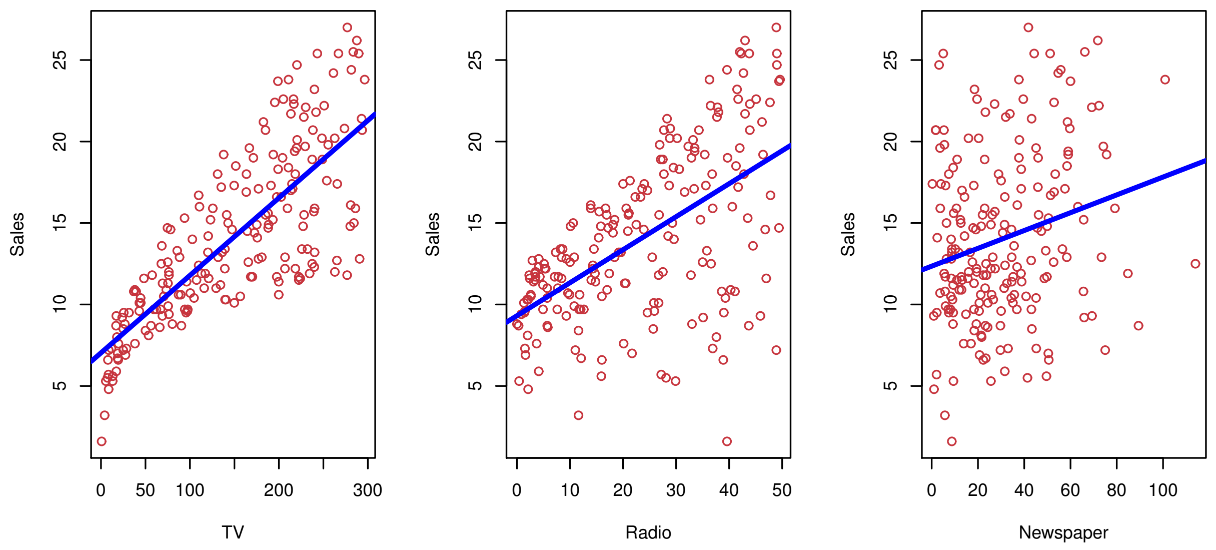


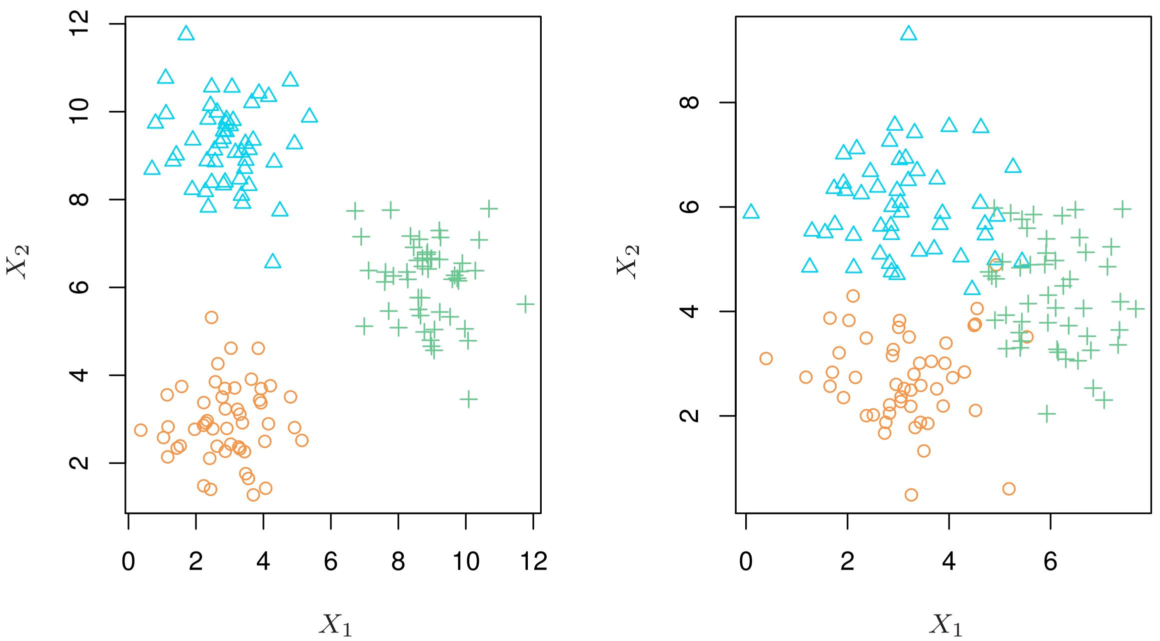
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Supervised vs. unsupervised learning

supervised learning

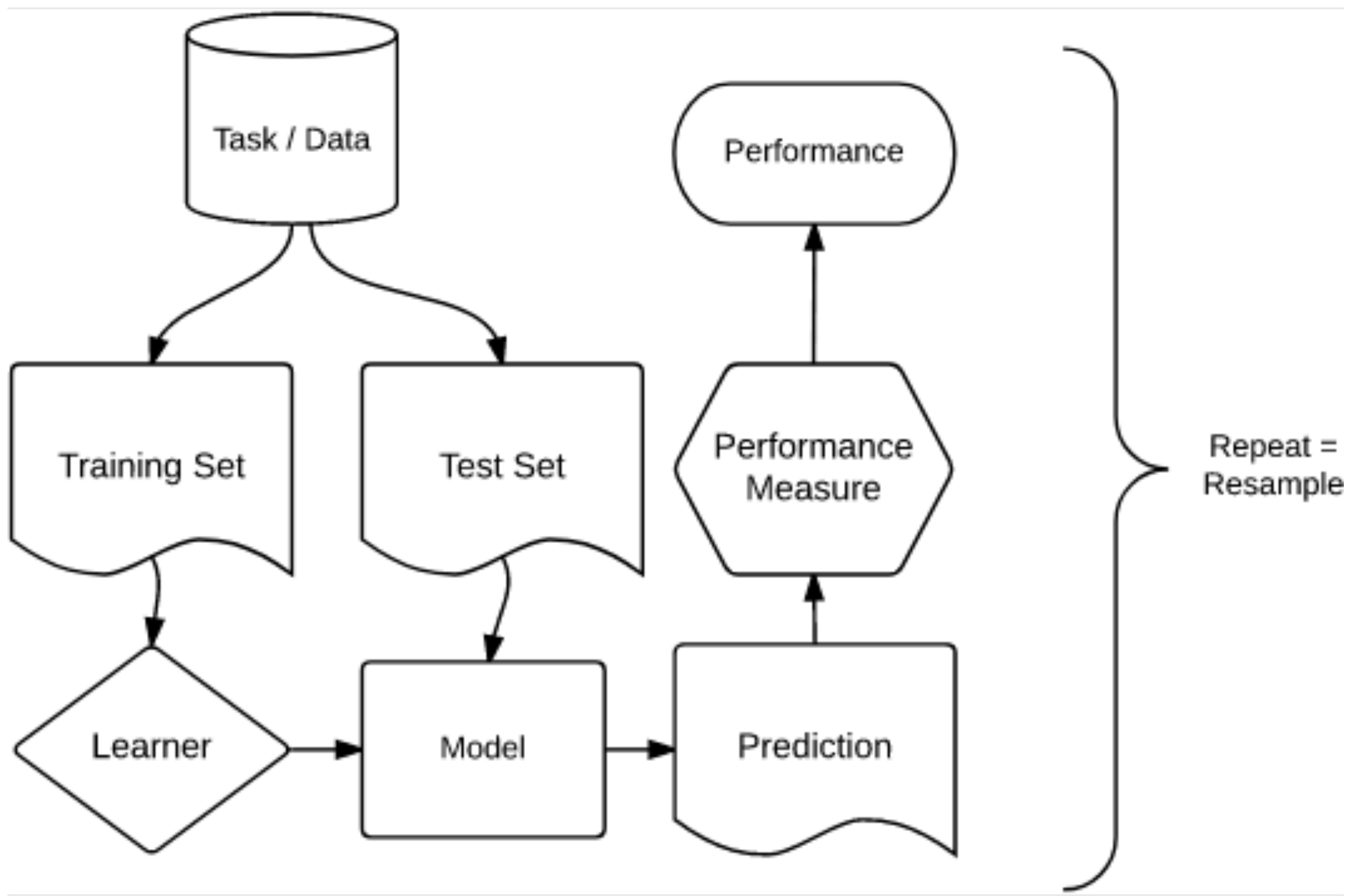
unsupervised learning





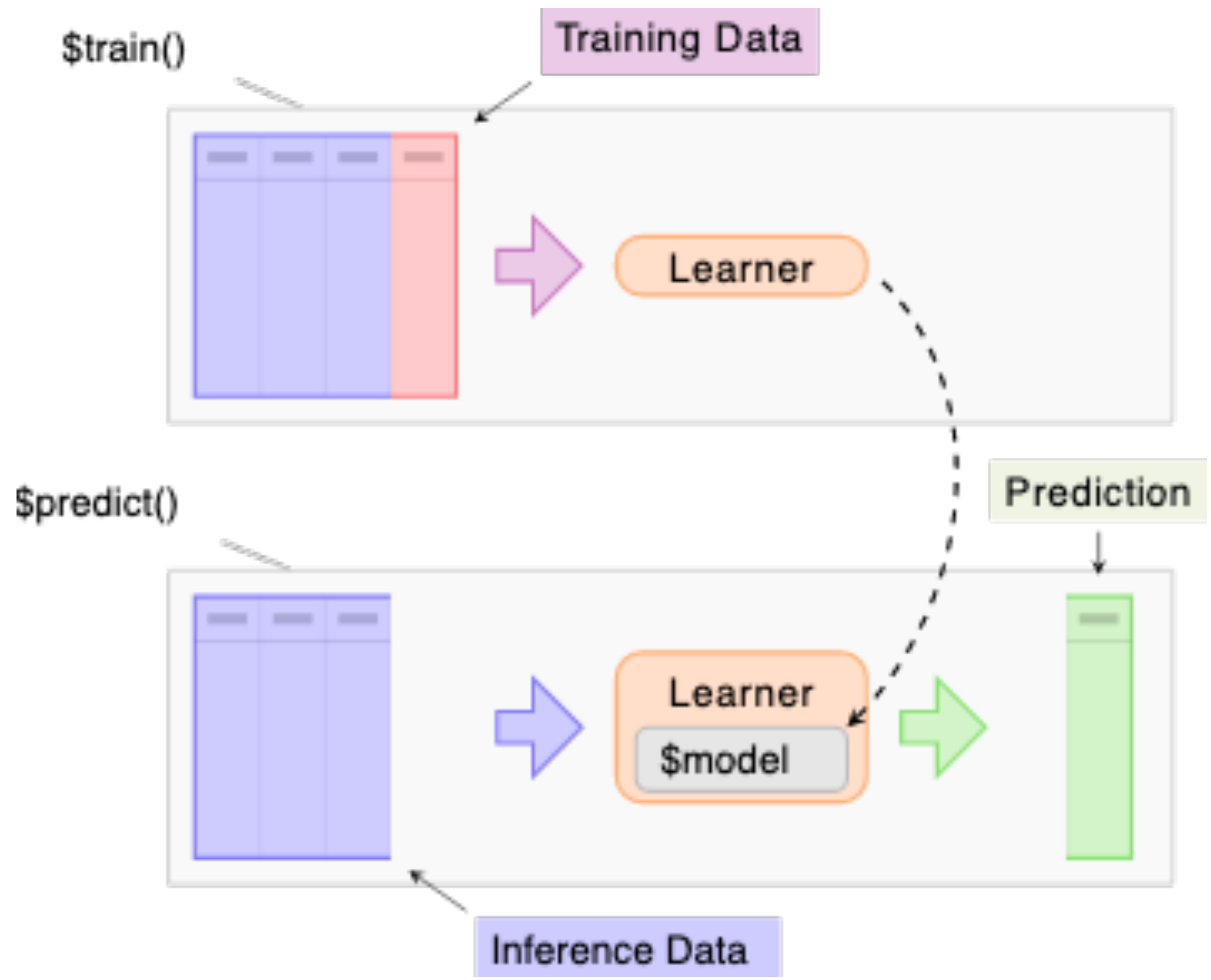
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Typical machine learning workflow



Lang, M., Schratz, P., Binder, M., Pfisterer, F., Richter, J., Reich, N. G., & Bischl, B. (2020). *Mlr3 book*. https://mlr3book.mlr-org.com

Learn and predict on new data



Lang, M., Schratz, P., Binder, M., Pfisterer, F., Richter, J., Reich, N. G., & Bischl, B. (2020). *Mlr3 book*. https://mlr3book.mlr-org.com

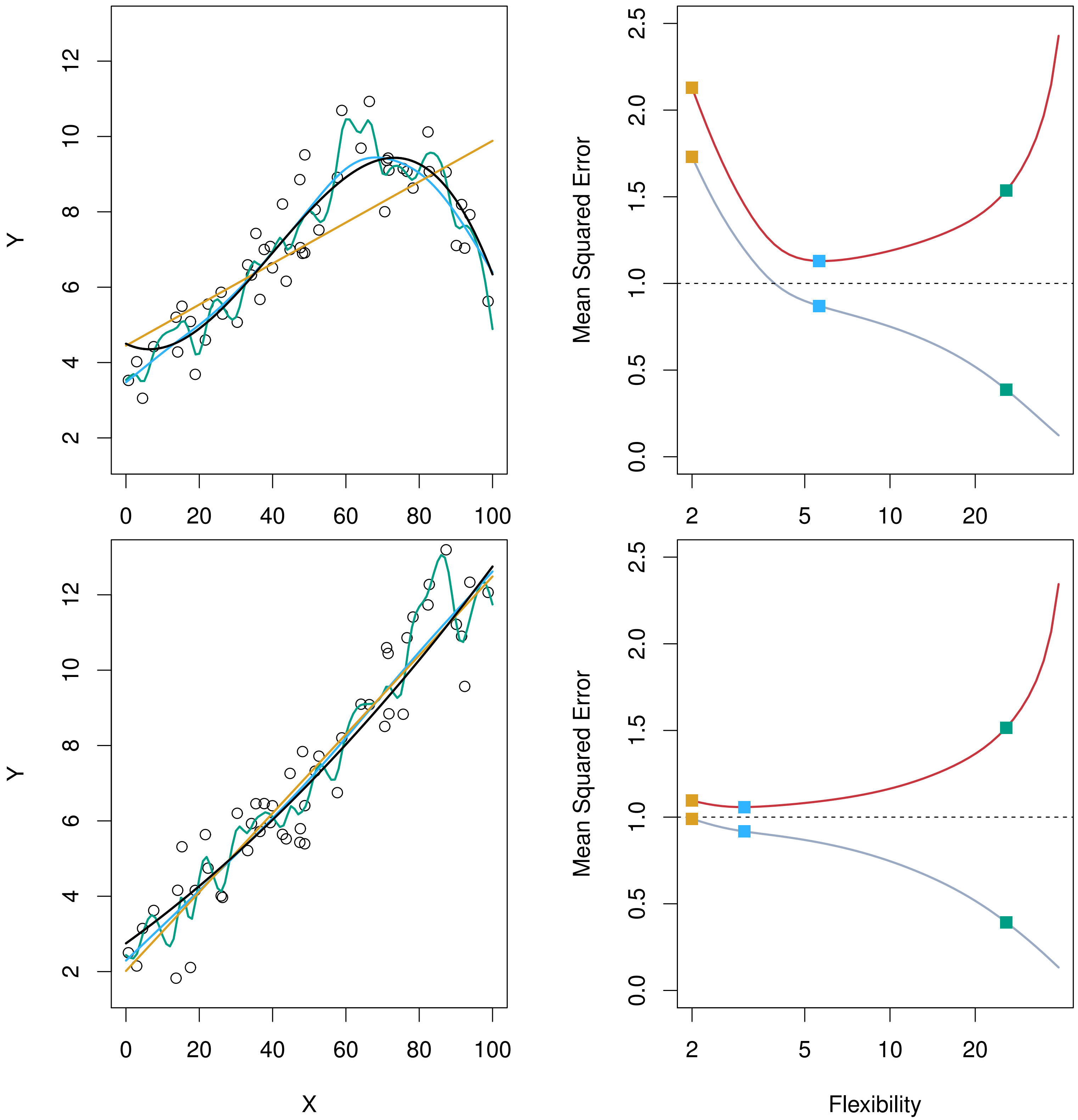
Model evaluation

Mean Square Error (MSE)

minimize $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2$$

$$MSE_{Test} = \frac{1}{n}\sum(y_0 - \hat{x}_o))^2$$

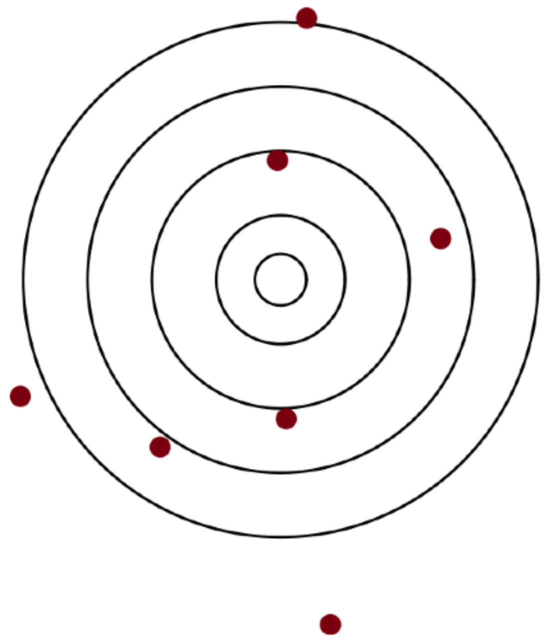
MSE Train ≠ MSE Test



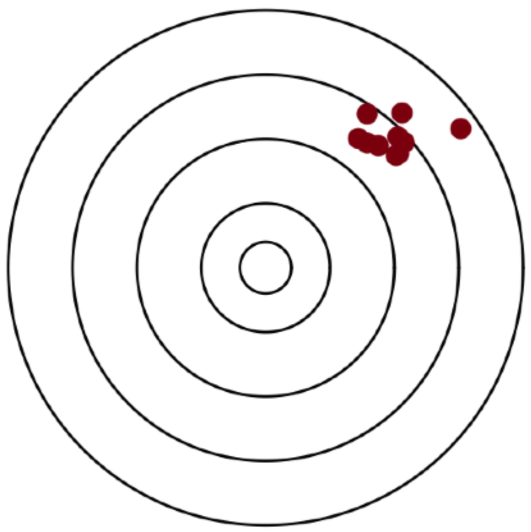
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Variance vs. bias

variance

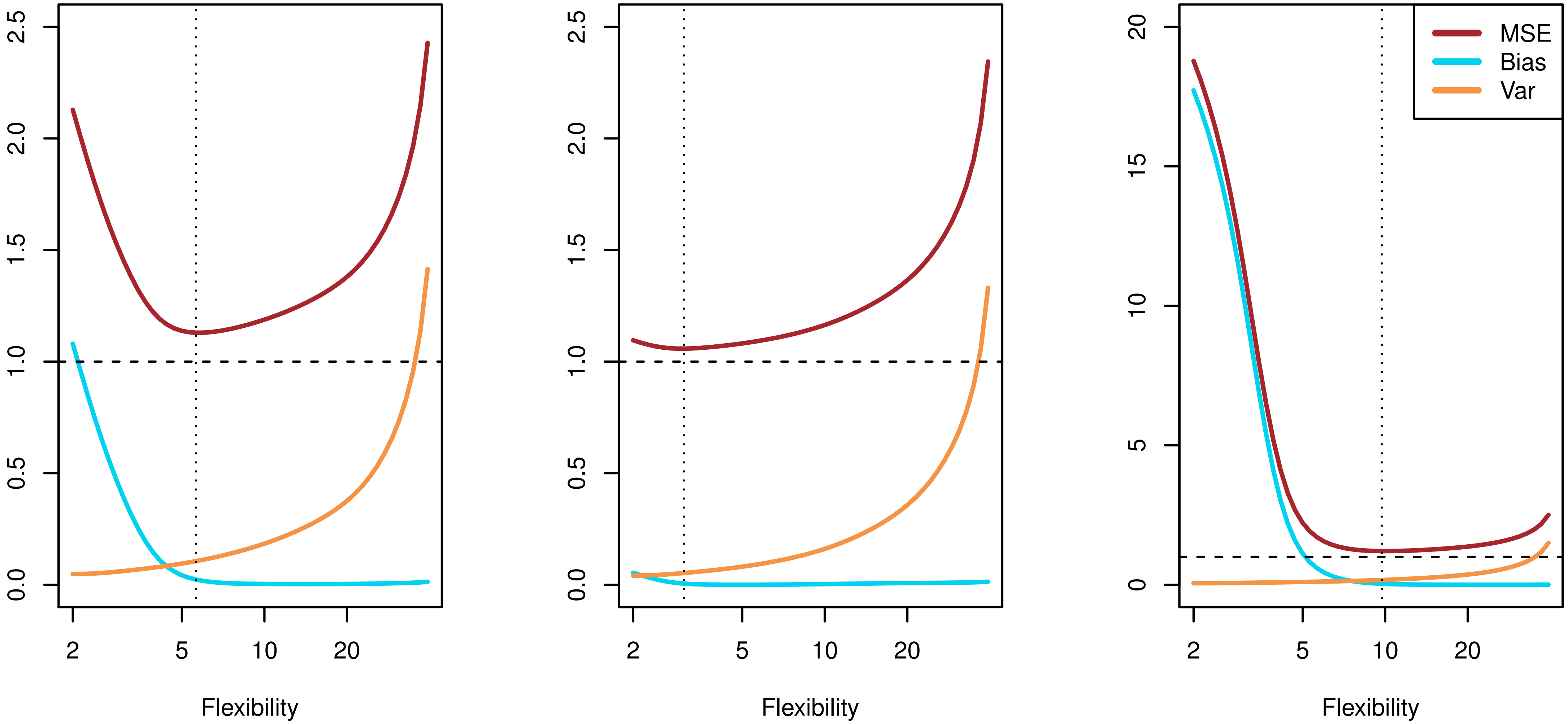


bias



MSE = Bias + Variance

f	non-linear, high error	nearly linear	non-linear, small error

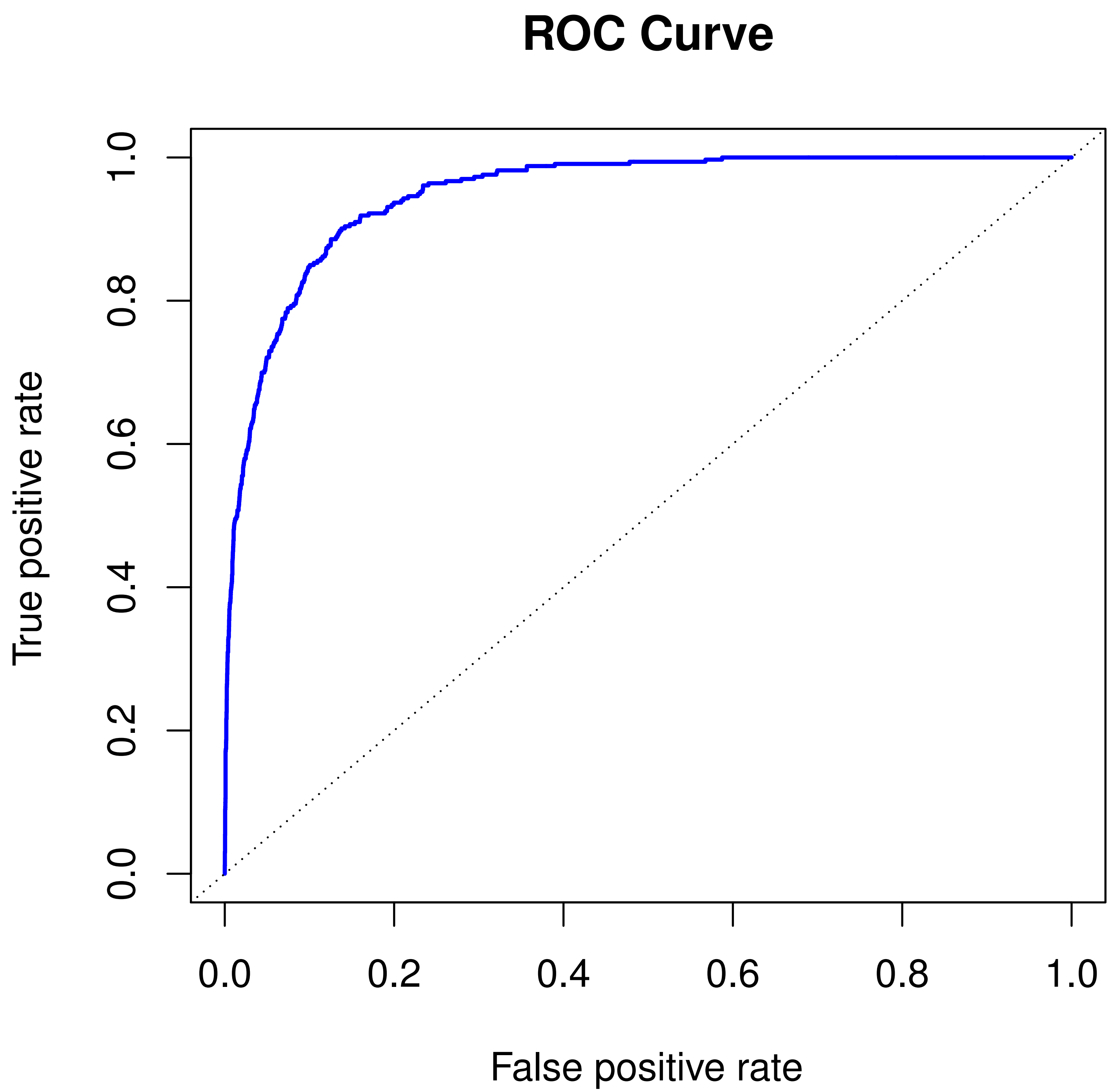


James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Model evaluation in classification models

$$e = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

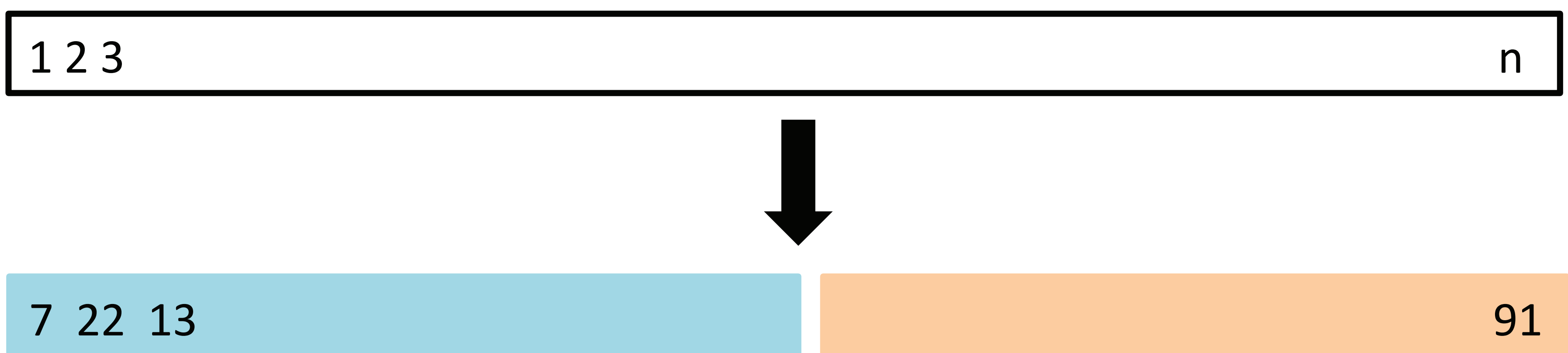
Klassifikationsmodelle visuell beurteilen



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

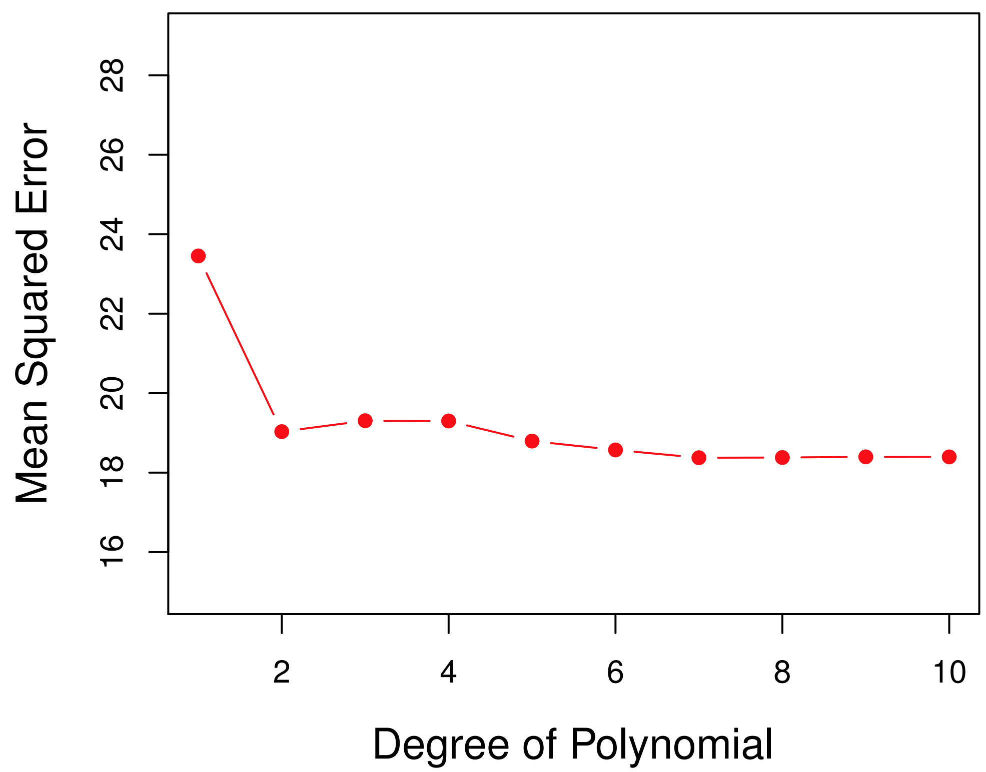
Resampling

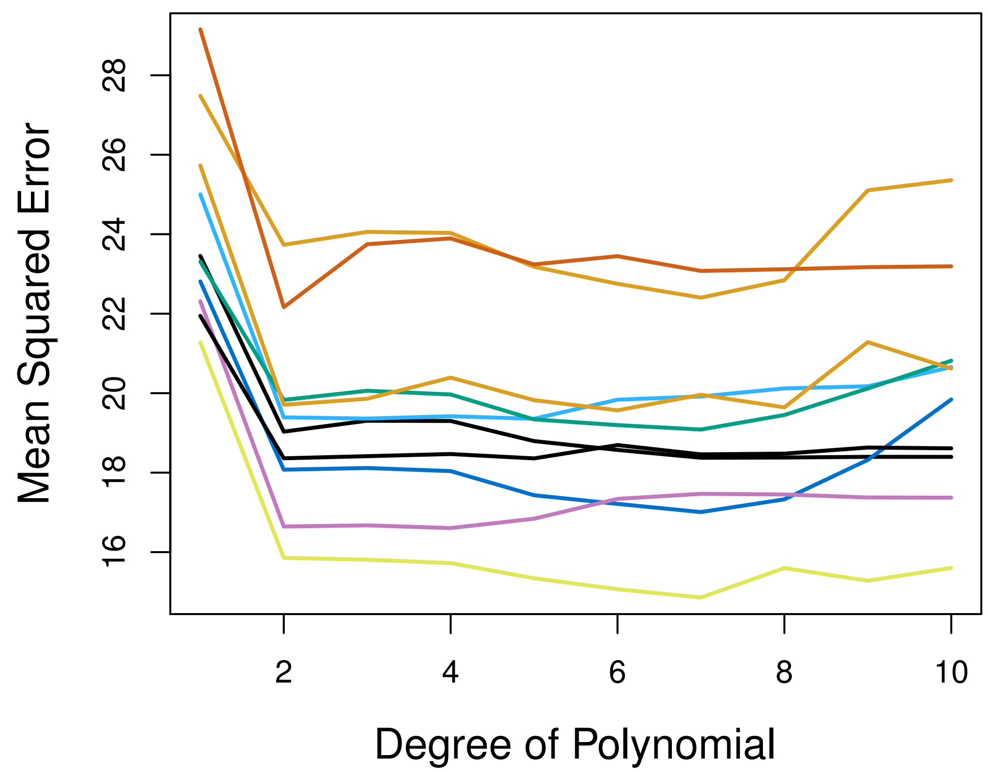
Train-Test split



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

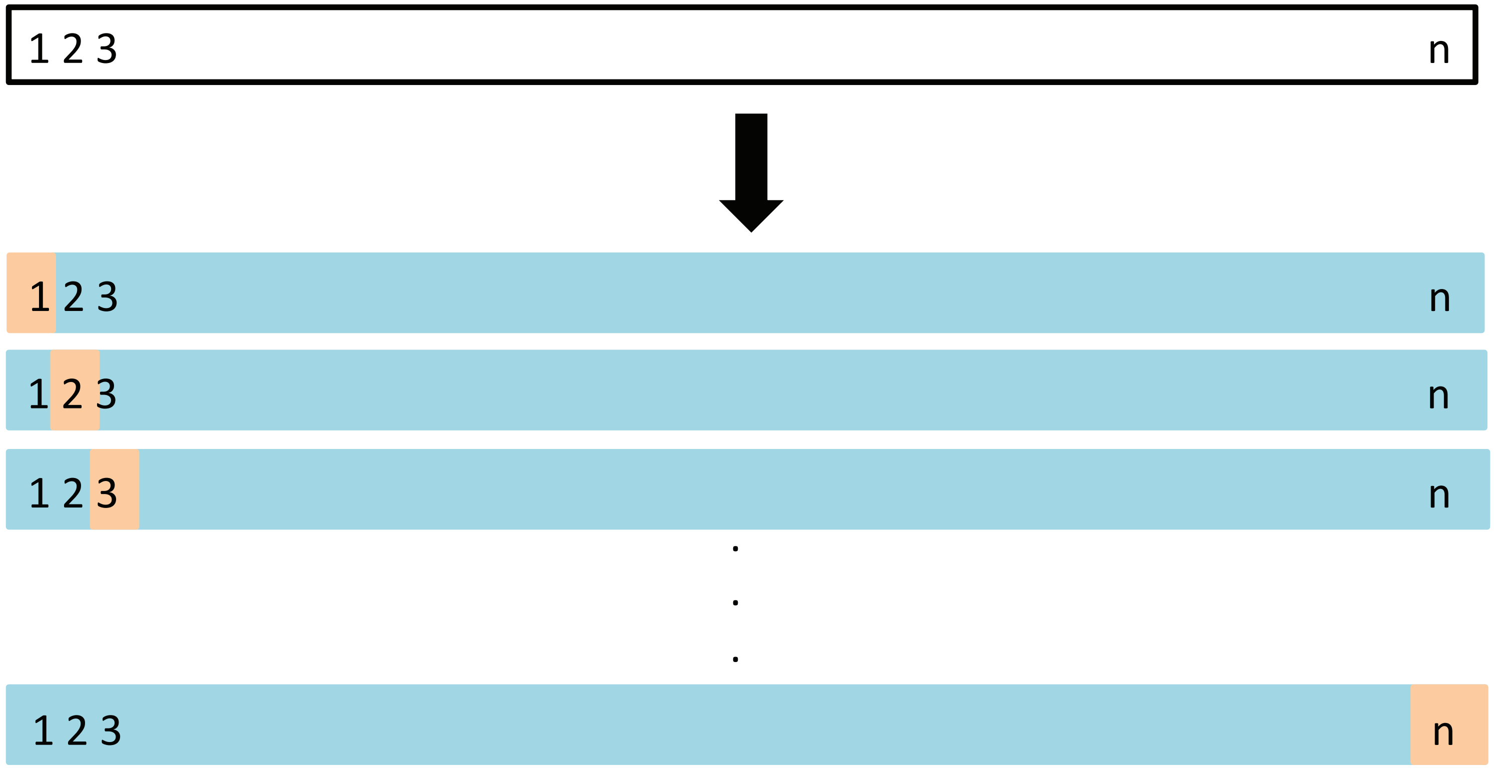
Train-Test split yields highly variable Test MSE





James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

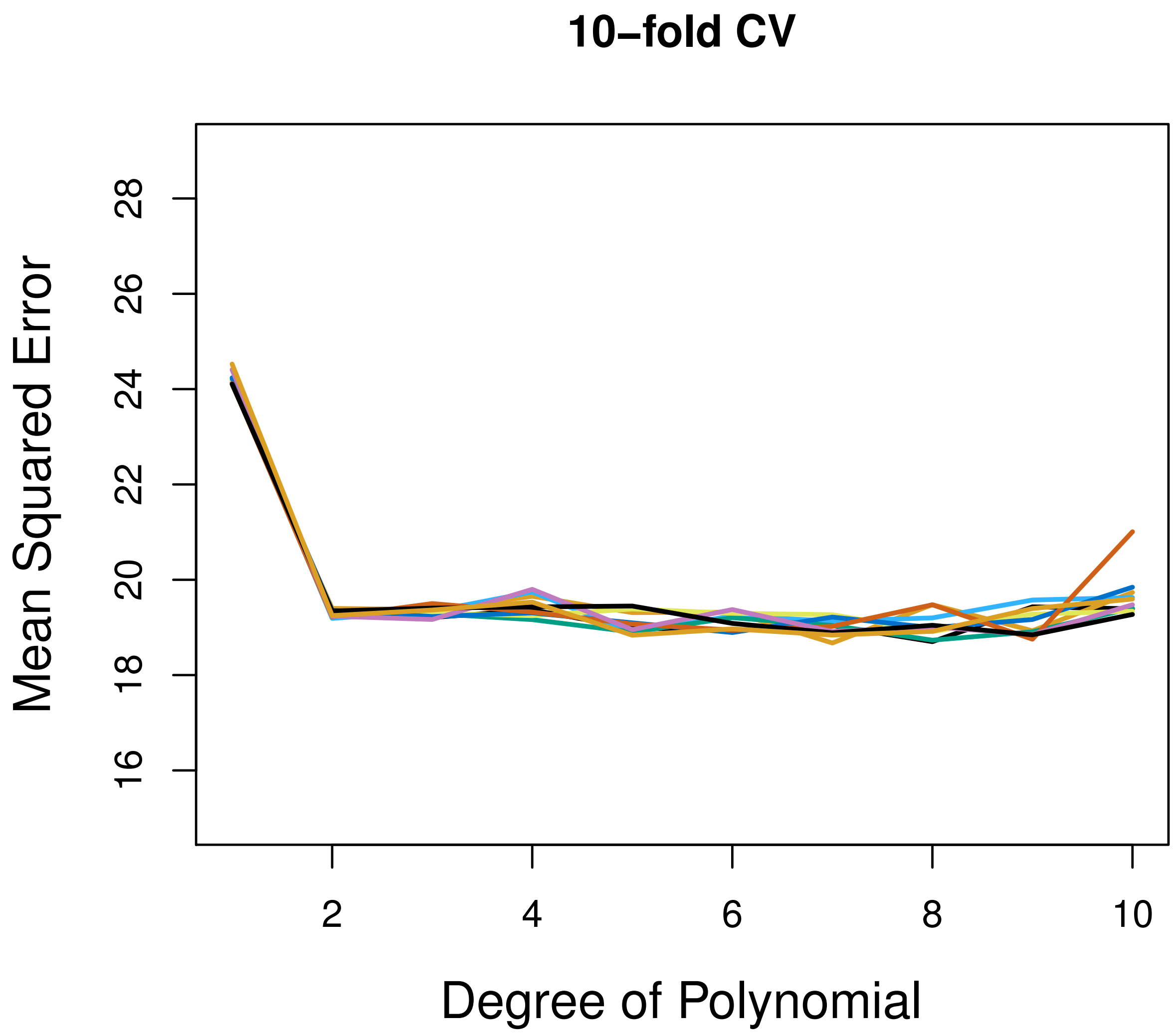
k-fold cross-validation (k-fold cc)



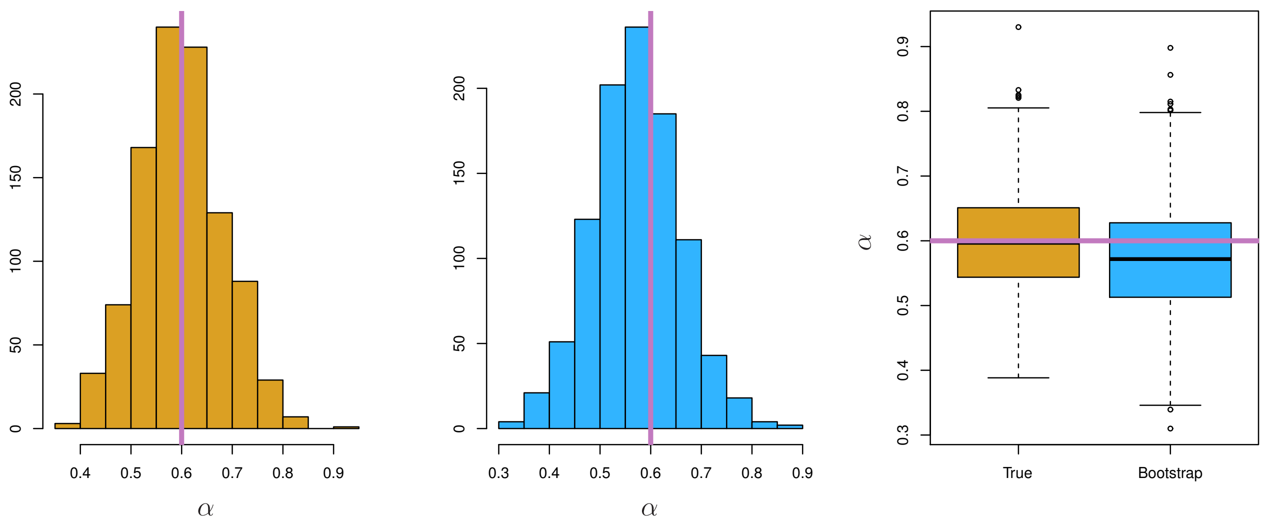
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

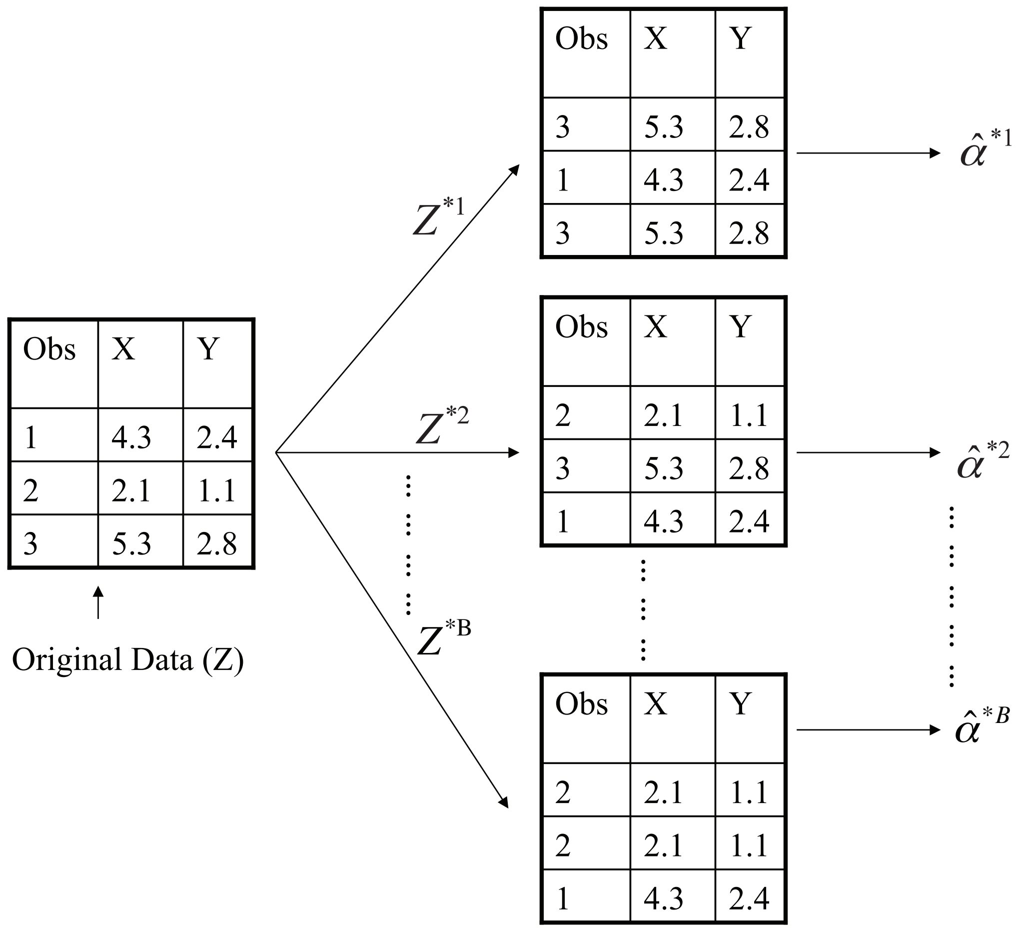
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

k-fold cross-validation (k-fold cc)



Bootstrap

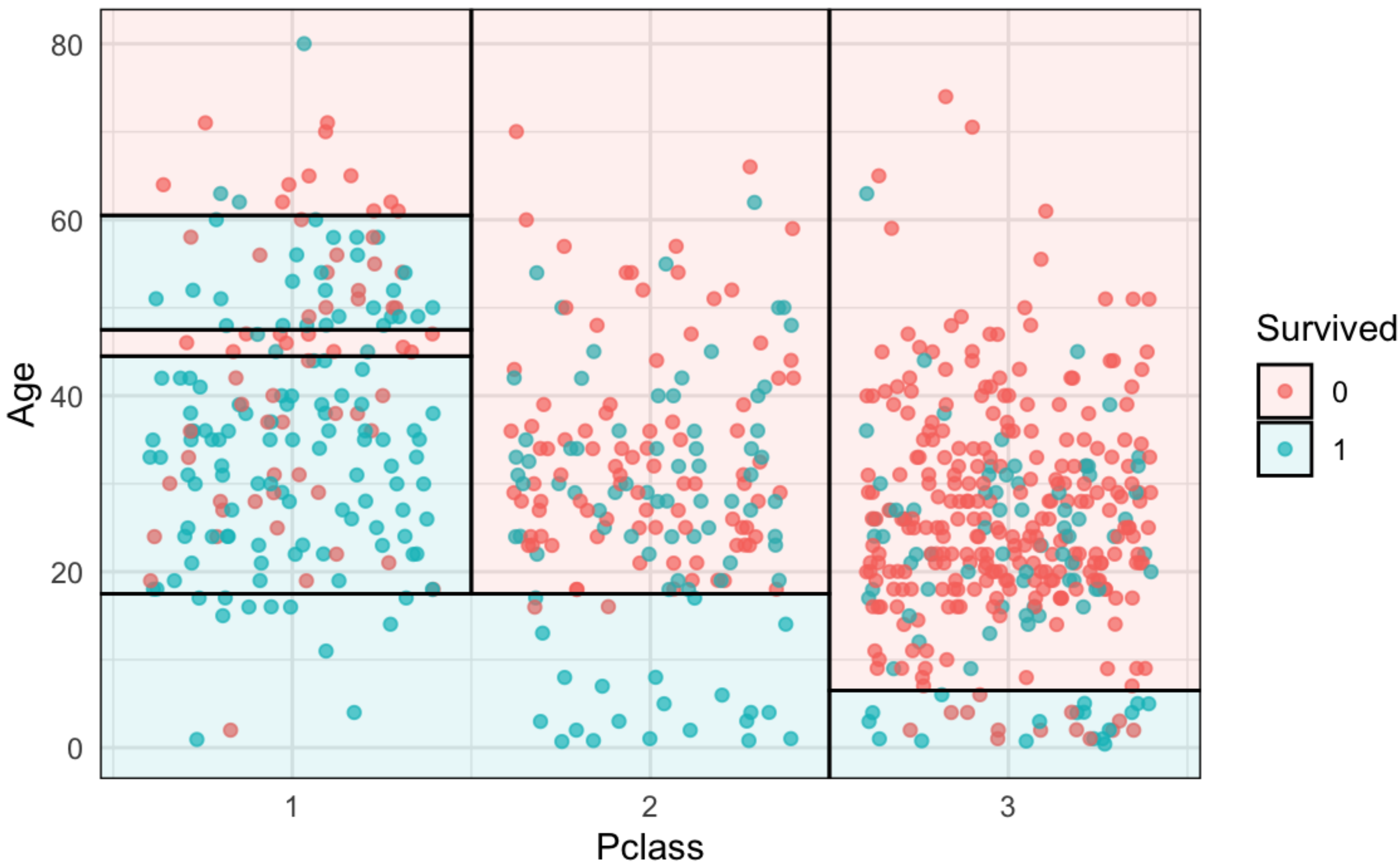




James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

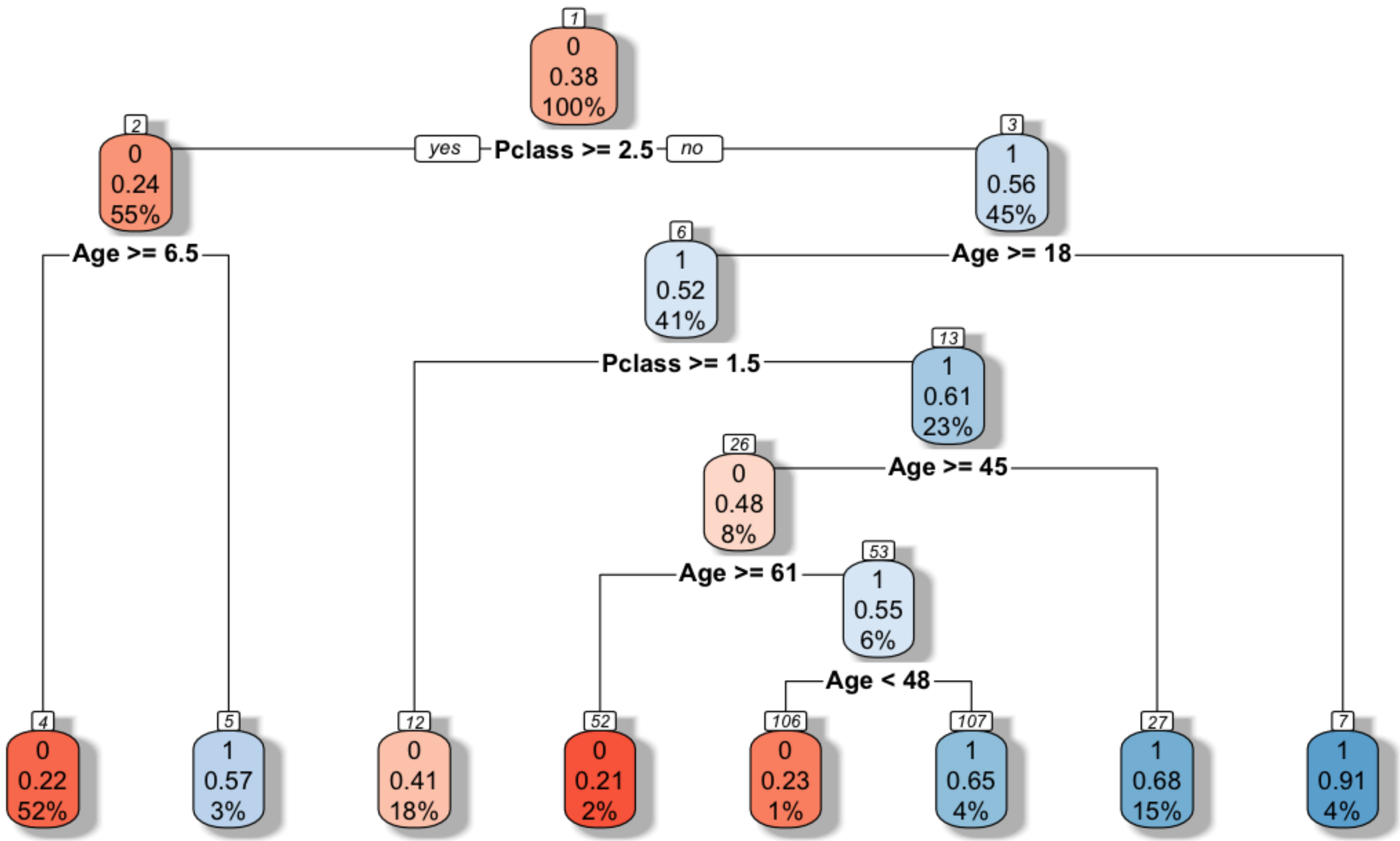
Tree-based Methods

Minimize the error, be pure



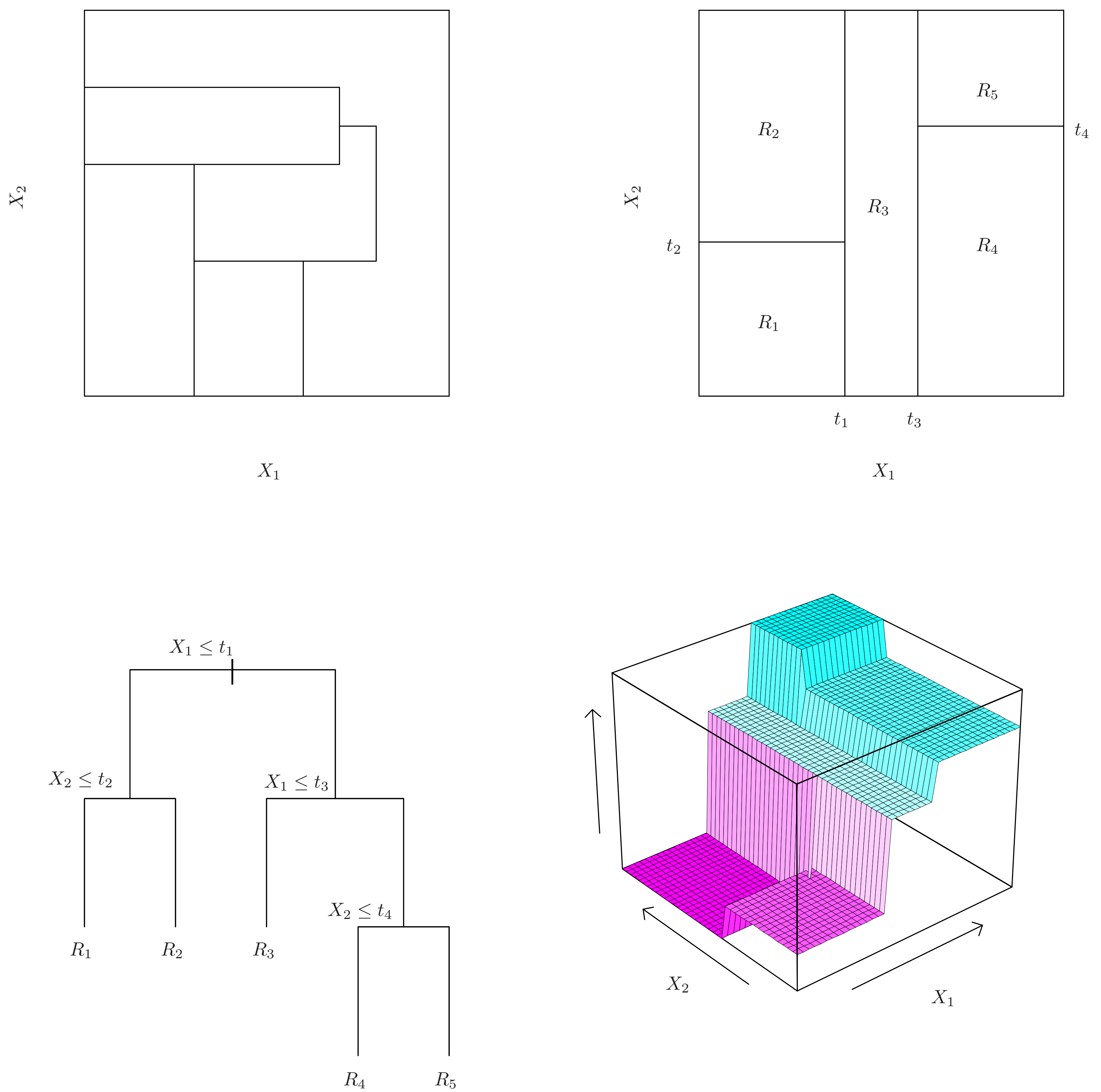
https://data-se.netlify.app/2020/10/17/visualizing-decision-trees/

Many Yes-No decisions yield a tree



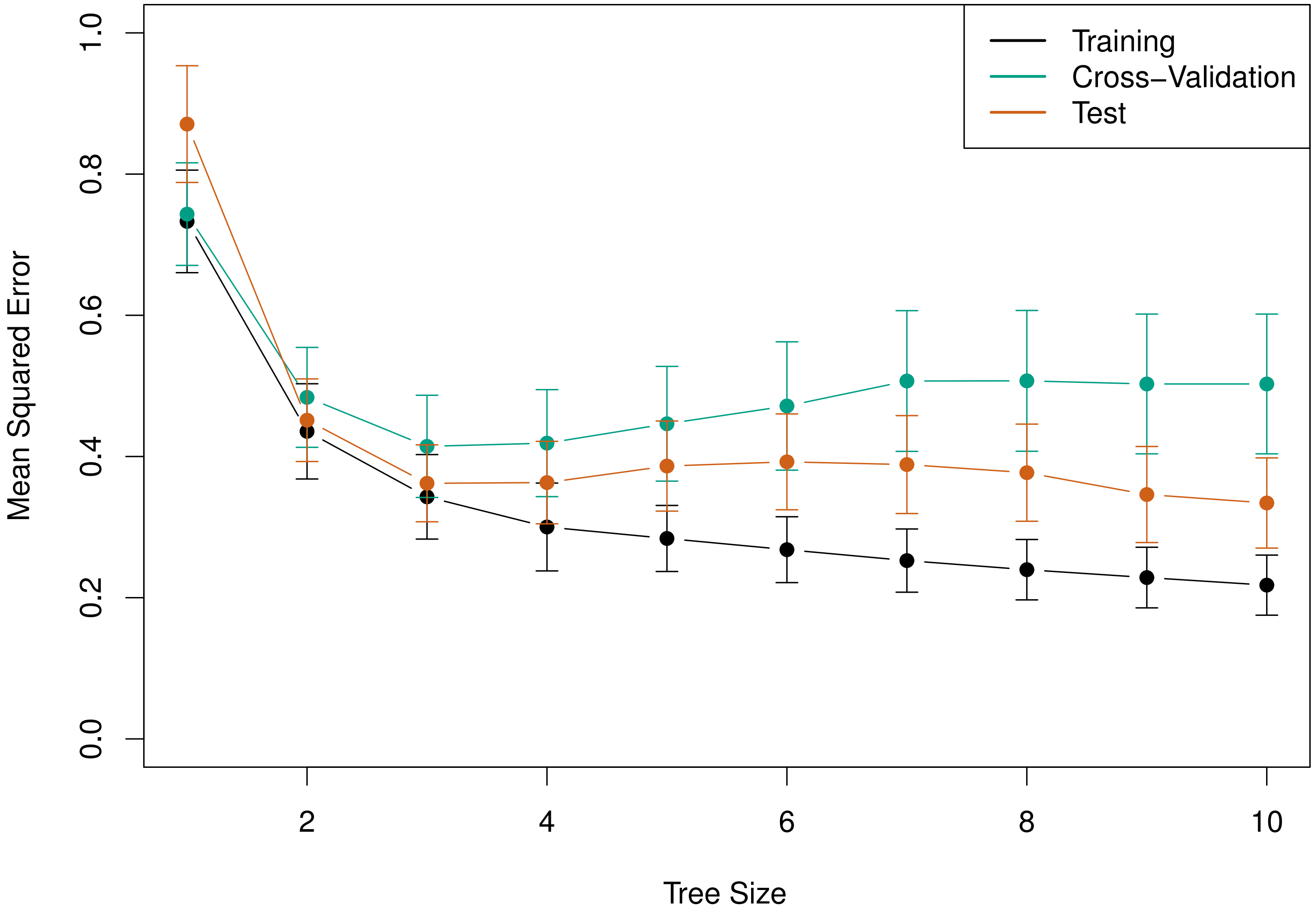
https://data-se.netlify.app/2020/10/17/visualizing-decision-trees/

Different visualisations of a tree



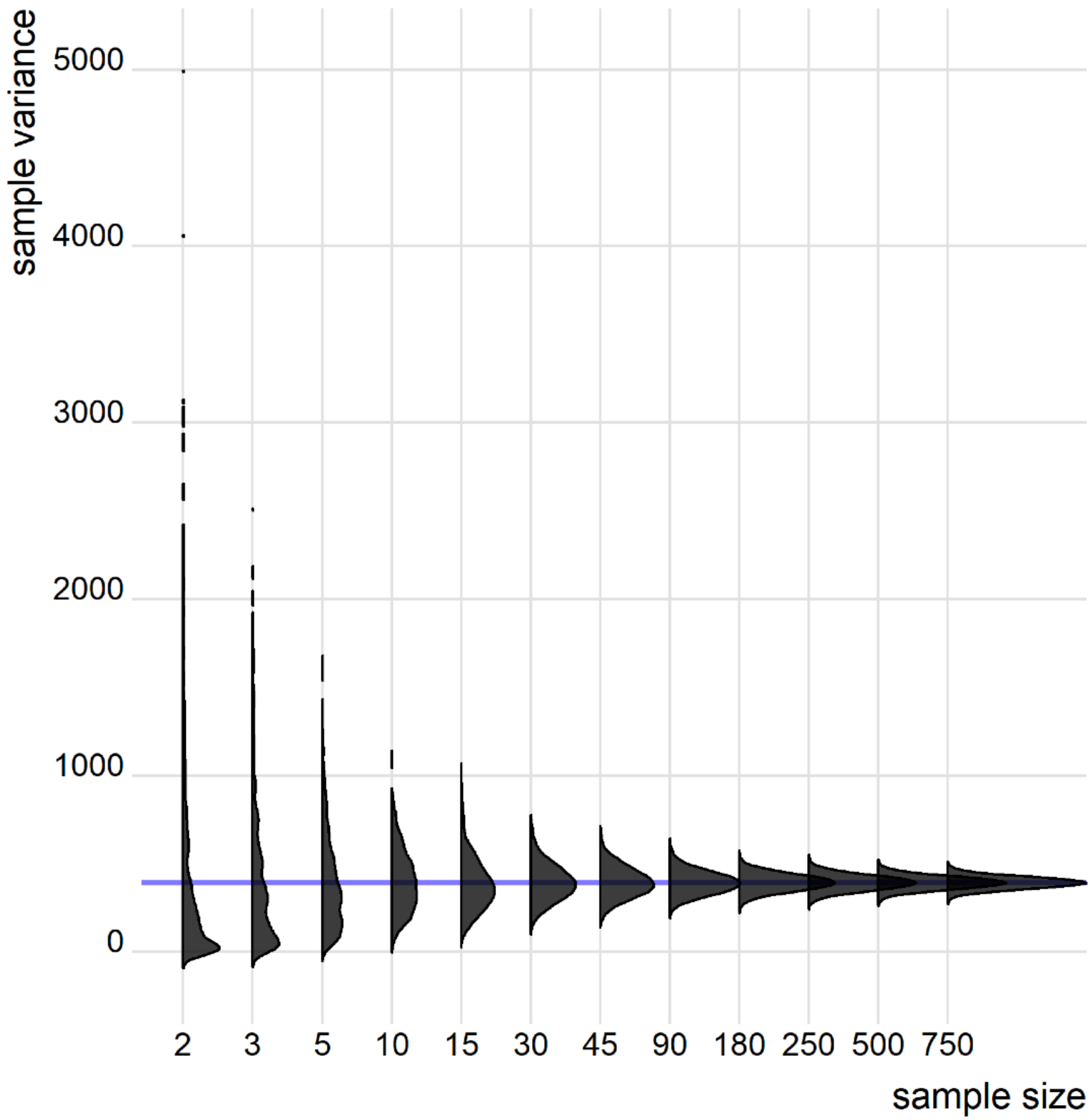
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Bigger trees, better prediction?



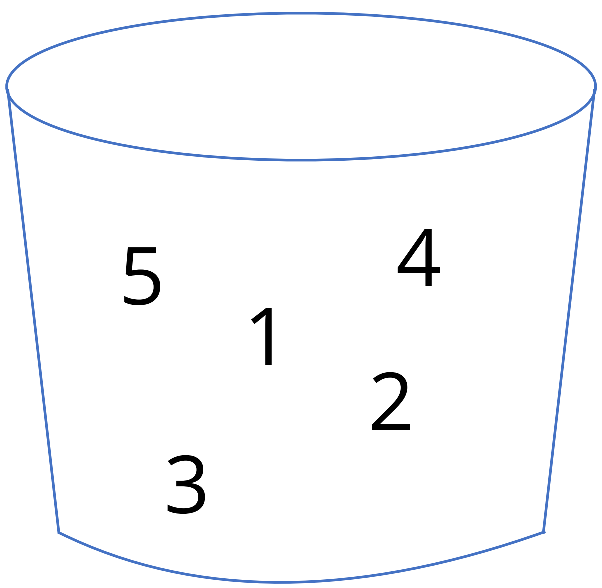
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

More trees yield a less variable estimate



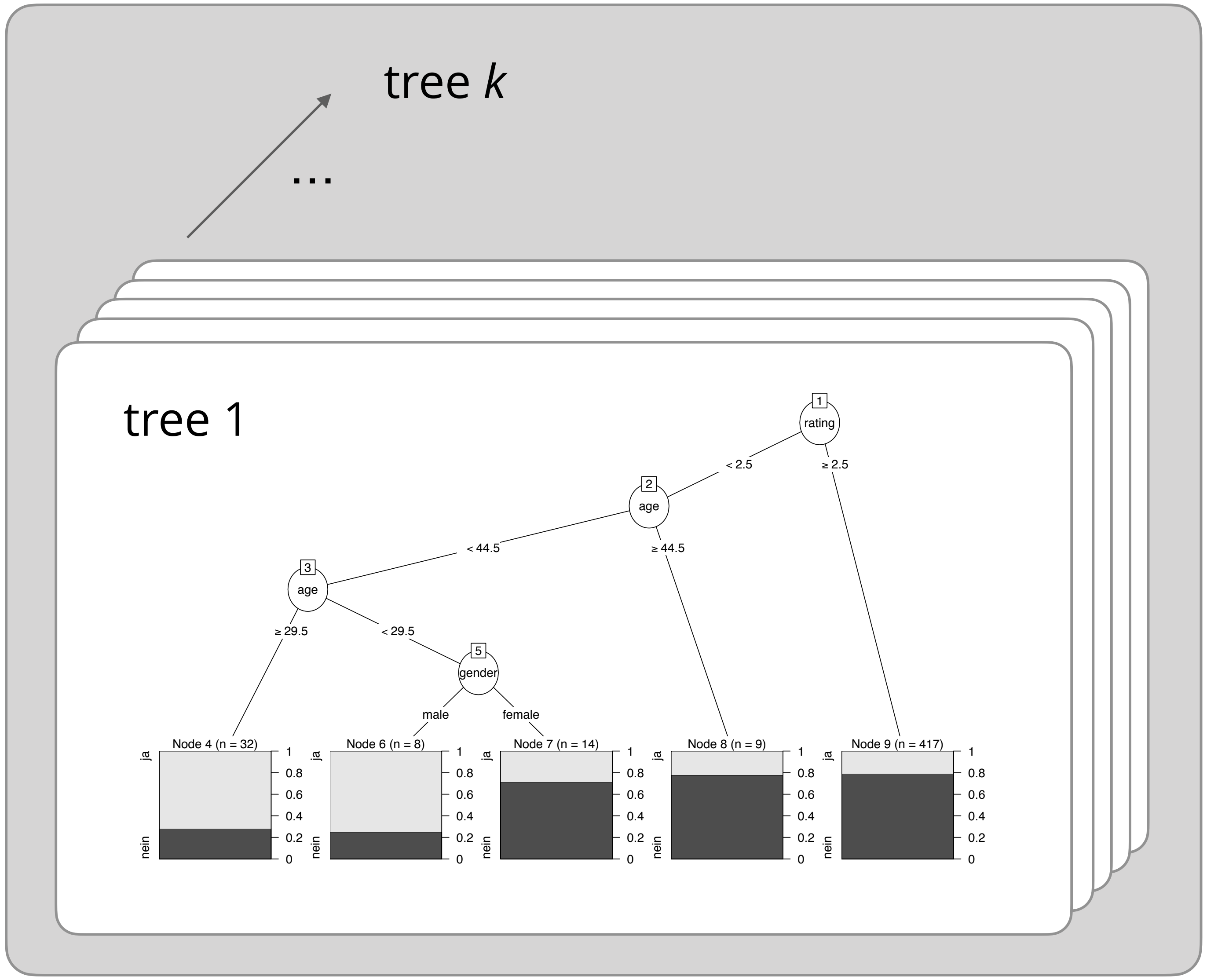
https://vbaliga.github.io/sample-variance-distributions-ggridges/

Bootstrapping



in bag	out of bag
1 1 3 2 2	4 5
2 1 4 2 2	3 5
5 1 5 2 5	3 4
...	...

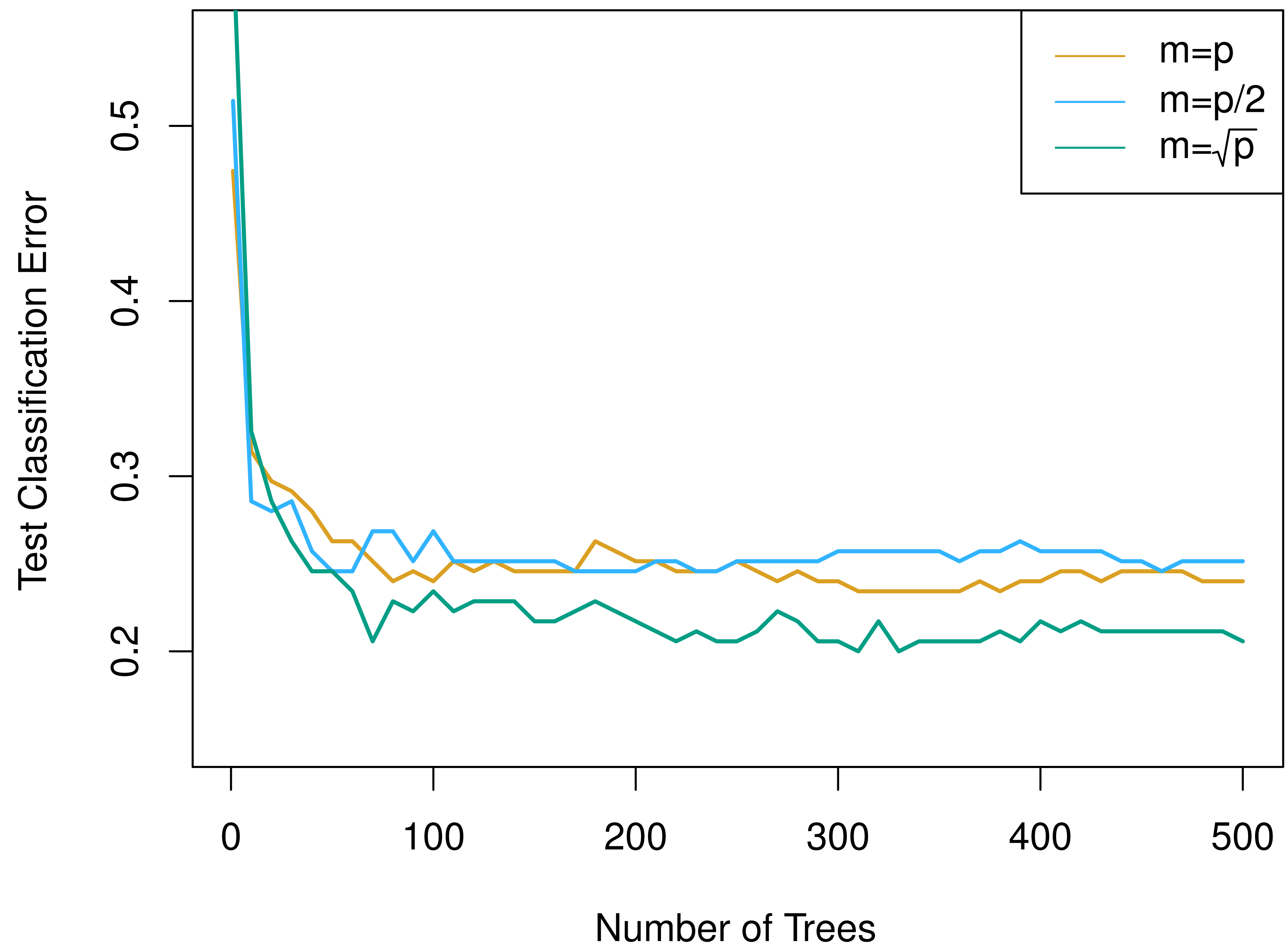
Random Forests



Random Forest

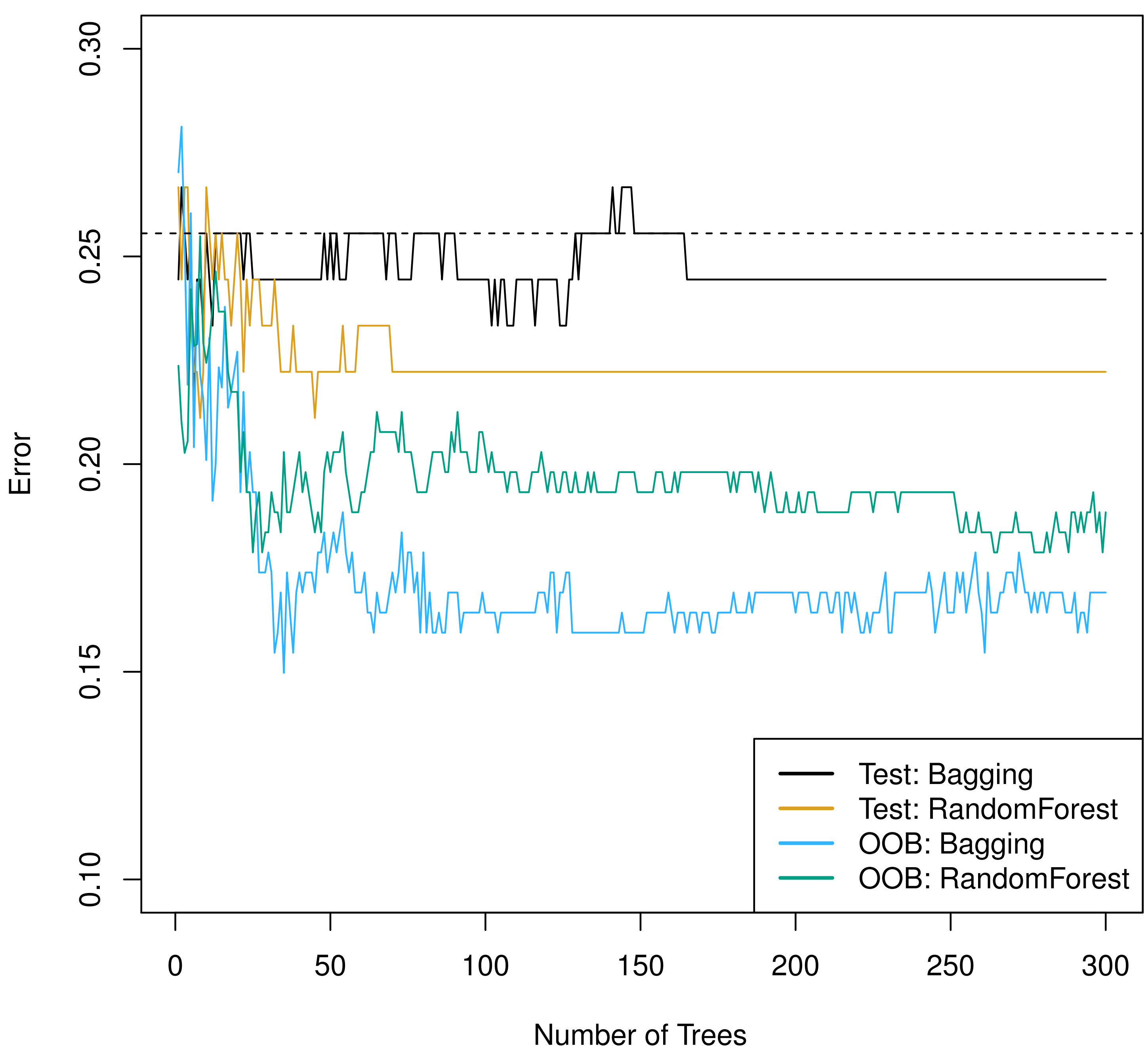
- random sample of predictors at each split

Comparing different random forest models



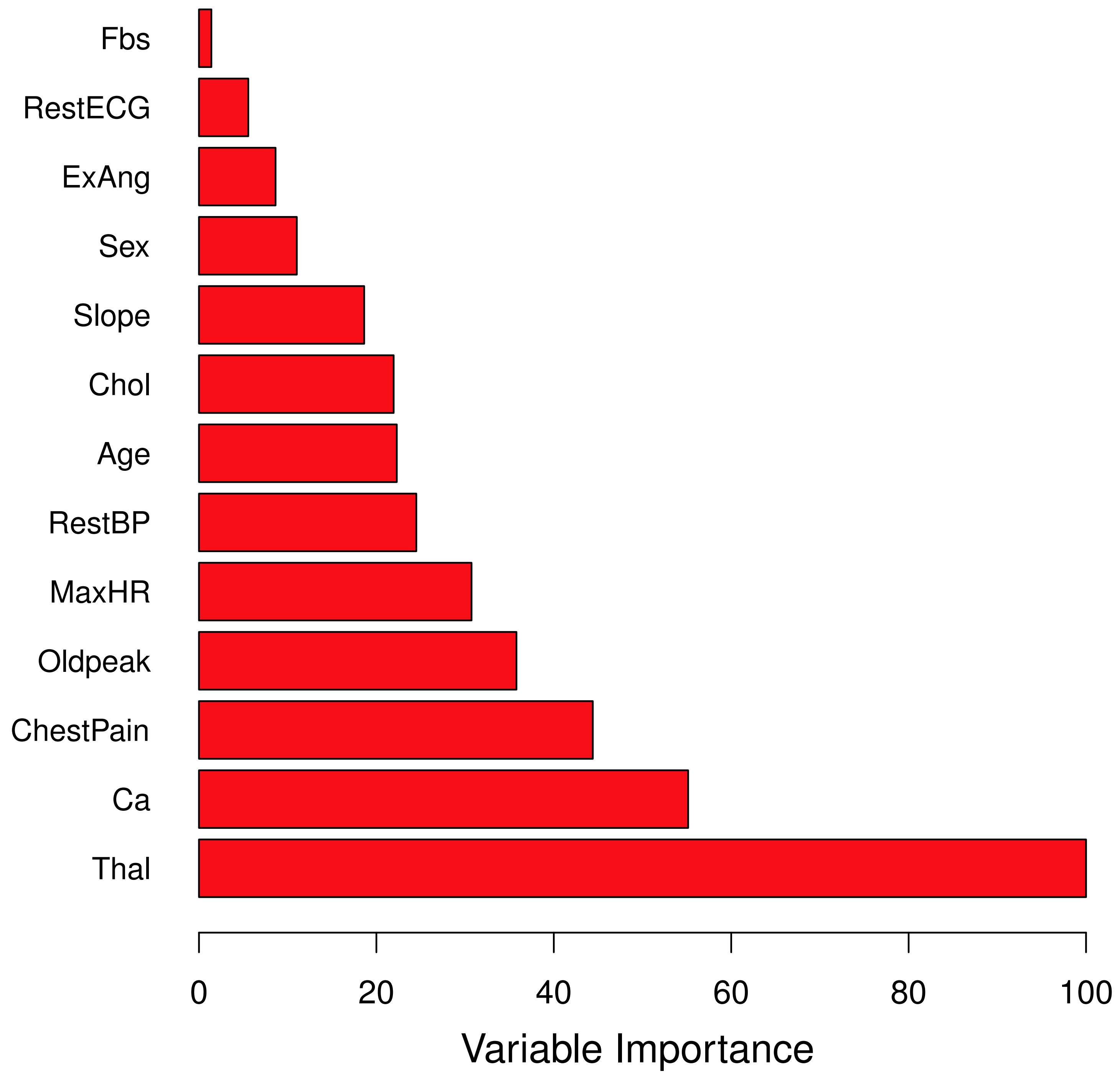
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Comparing trees, bagging and random forests



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

Variable importance



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf