MACHINE LEARNING TOOLBOX

Welcome to the
Machine Learning Toolbox!

Supervised learning

- `caret` R package

- Automates *supervised learning* (a.k.a. *predictive modeling*)

- Target variable

Supervised learning

- Two types of predictive models

    - Classification ⟶ **Qualitative**

    - Regression ⟶ **Quantitative**

- Use *metrics* to evaluate models

    - Quantifiable

    - Objective

- *Root Mean Squared Error* (RMSE) for regression (e.g. `lm()`)

Evaluating model performance

- Common to calculate in-sample RMSE

  - Too optimistic

  - Leads to overfitting

- Better to calculate out-of-sample error (a la `caret`)

  - Simulates real-world usage

  - Helps avoid overfitting

In-sample error

```
> # Fit a model to the mtcars data
> data(mtcars)
> model <- lm(mpg ~ hp, mtcars[1:20, ])

> # Predict in-sample
> predicted <- predict(model, mtcars[1:20, ], type = "response")

> # Calculate RMSE
> actual <- mtcars[1:20, "mpg"]
> sqrt(mean((predicted - actual)^2))
[1] 3.172132
```

The Machine Learning Toolbox

Let's practice!

MACHINE LEARNING TOOLBOX

Out-of-sample
error measures

Out-of-sample error

- Want models that don't overfit and generalize well

- Do the models perform well on <u>new</u> data?

- Test models on new data, or a *test set*

  - Key insight of machine learning

  - In-sample validation almost guarantees overfitting

- Primary goal of `caret` and this course: don't overfit

Example: out-of-sample RMSE

```
> # Fit a model to the mtcars data
> data(mtcars)
> model <- lm(mpg ~ hp, mtcars[1:20, ])

> # Predict out-of-sample
> predicted <- predict(model, mtcars[21:32, ], type = "response")

> # Evaluate error
> actual <- mtcars[21:32, "mpg"]
> sqrt(mean((predicted - actual)^2))
[1] 5.507236
```

Alternatives:
createResamples()
createFolds()

Compare to in-sample RMSE

```r
> # Fit a model to the full dataset
> model2 <- lm(mpg ~ hp, mtcars)

> # Predict in-sample
> predicted2 <- predict(model, mtcars, type = "response")

> # Evaluate error
> actual2 <- mtcars[, "mpg"]
> sqrt(mean((predicted2 - actual2)^2))
[1] 3.74     Compare to out-of-sample RMSE of 5.5
```

MACHINE LEARNING TOOLBOX

Let's practice!

Cross-validation

Cross-validation

Fit final model on full dataset

Full dataset $\rightarrow$ Final model

CV is 11x as
expensive as fitting
a single model!
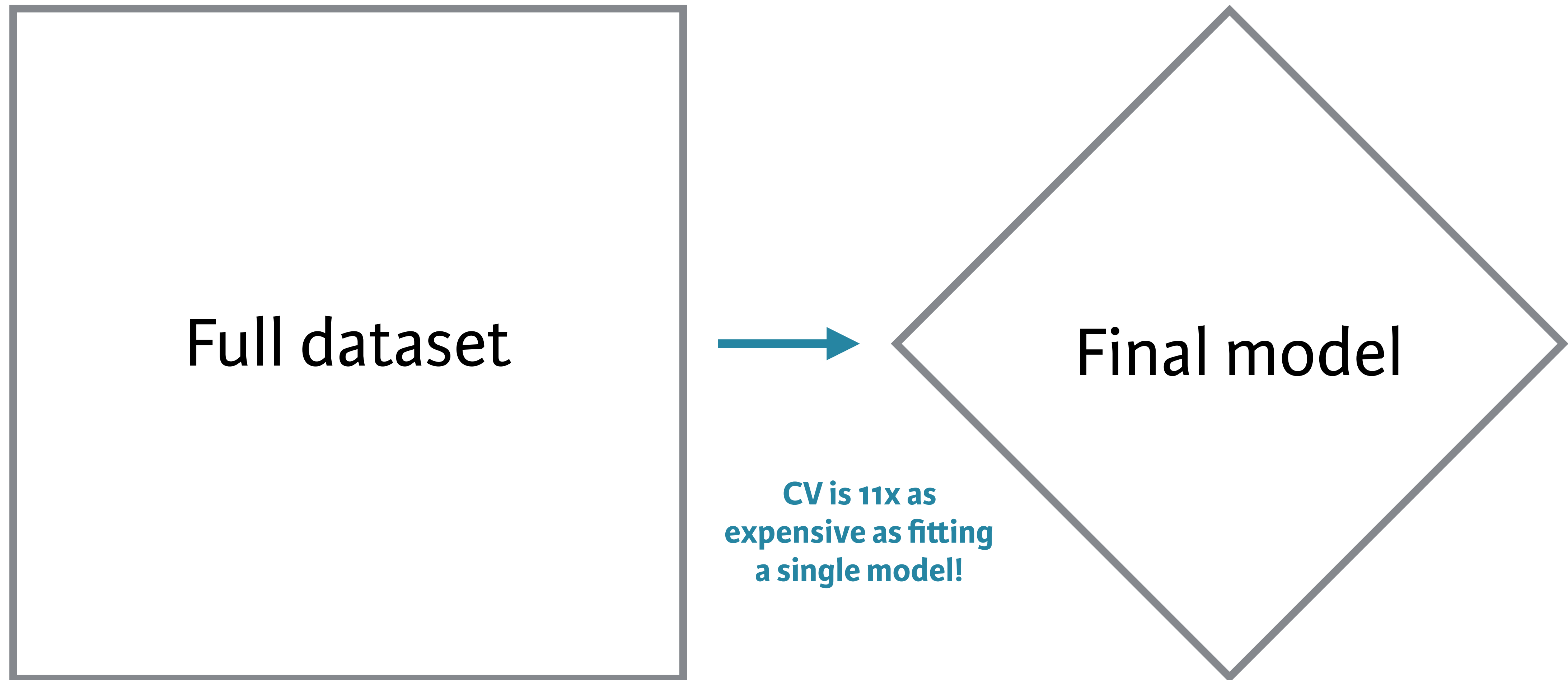
Cross-validation

```
> # Set seed for reproducibility
> library(caret)
> data(mtcars)
> set.seed(42)

> # Fit linear regression model
> model <- train(mpg ~ hp, mtcars,
                method = "lm",
                trControl = trainControl(
                  method = "cv", number = 10,
                  verboseIter = TRUE
                )
  )
+ Fold01: parameter=none
+ Fold02: parameter=none
        ...
- Fold10: parameter=none
Aggregating results
Fitting final model on full training set
```

MACHINE LEARNING TOOLBOX

Let's practice!