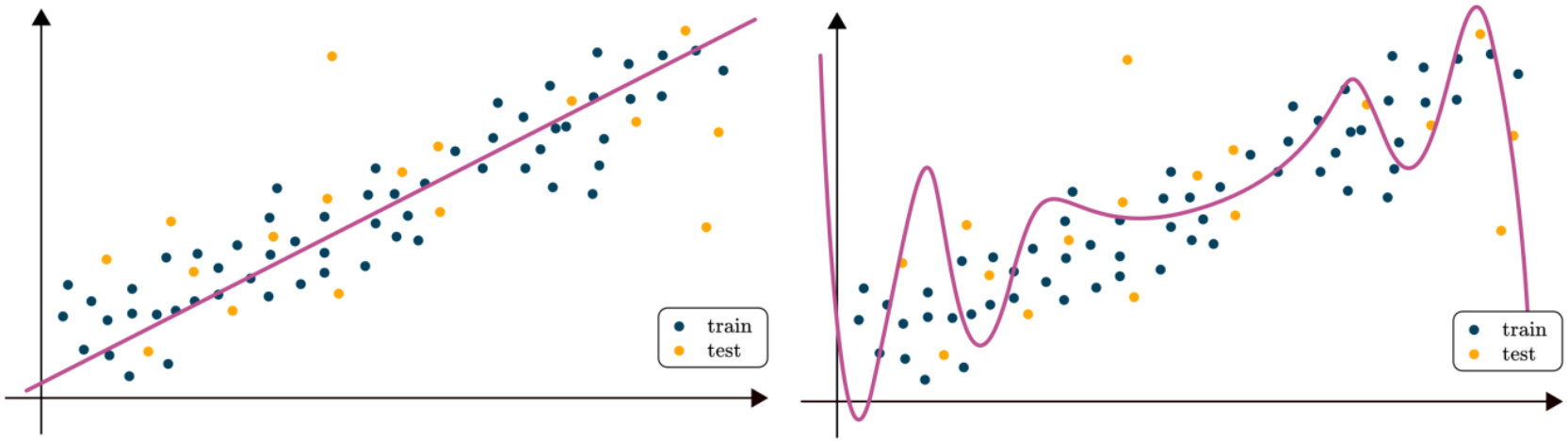
Generalizability

PSY-GS 8875 Behavioral Data Science



Overview

Overview: Week 3

Readings

- ESL Chapters: 7.10 and 7.11
- HML: Chapter 2
- Yarkoni - 2022

Optional

- Shen et al. - 2017

- Data splitting (training/testing/validation)
- Bootstrap
- Relative importance
- k-folds cross-validation
- Activity: model evaluation of math success

Generalizability

Generalizability

Consider the following scenario...

Scenario 1

Researchers fit a linear model with **X** predictors and **y** outcome. Other researchers, in a new sample, fit a linear model with **X** predictors and **y** outcome. The same **X** variables were (non-significant) in both models predicting **y**.

To what extent can we determine whether the linear model generalizes?

Generalizability

Consider the following scenario...

Scenario 1

Researchers fit a linear model with **X** predictors and **y** outcome. Other researchers, in a new sample, fit a linear model with **X** predictors and **y** outcome. The same **X** variables were (non-significant) in both models predicting **y**.

To what extent can we determine whether the linear model generalizes?

How often do you see this scenario in your literature?

Generalizability

Consider the following scenario...

Scenario 2

Researchers fit a linear model with **X** predictors and **y** outcome. Other researchers, in a new sample, use the same regression coefficients as the original linear model of **X** predictors and to predict **y** outcome. The R^2 was around 0.10.

To what extent can we determine whether the linear model generalizes?

Generalizability

Consider the following scenario...

Scenario 2

Researchers fit a linear model with **X** predictors and **y** outcome. Other researchers, in a new sample, use the same regression coefficients as the original linear model of **X** predictors and to predict **y** outcome. The R^2 was around 0.10.

To what extent can we determine whether the linear model generalizes?

How often do you see this scenario in your literature?

Generalizability

Consider the following scenario...

Scenario 3

Researchers fit a linear model with **X** predictors and **y** outcome using 70% of their sample and test their model's predictions (using the regression coefficients from the model fit to the 70% of the sample) on the remaining 30%. They find that their model has an R^2 around 0.12 in the test dataset. Other researchers, in a new sample, use the same regression coefficients as the original linear model (using the 70% of the sample). Their R^2 was around 0.14.

To what extent can we determine whether the linear model generalizes?

Generalizability

Consider the following scenario...

Scenario 3

Researchers fit a linear model with **X** predictors and **y** outcome using 70% of their sample and test their model's predictions (using the regression coefficients from the model fit to the 70% of the sample) on the remaining 30%. They find that their model has an R^2 around 0.12 in the test dataset. Other researchers, in a new sample, use the same regression coefficients as the original linear model (using the 70% of the sample). Their R^2 was around 0.14.

To what extent can we determine whether the linear model generalizes?

How often do you see this scenario in your literature?

Generalizability

generalizability: extent to which a model can make predictions beyond the data it was fit

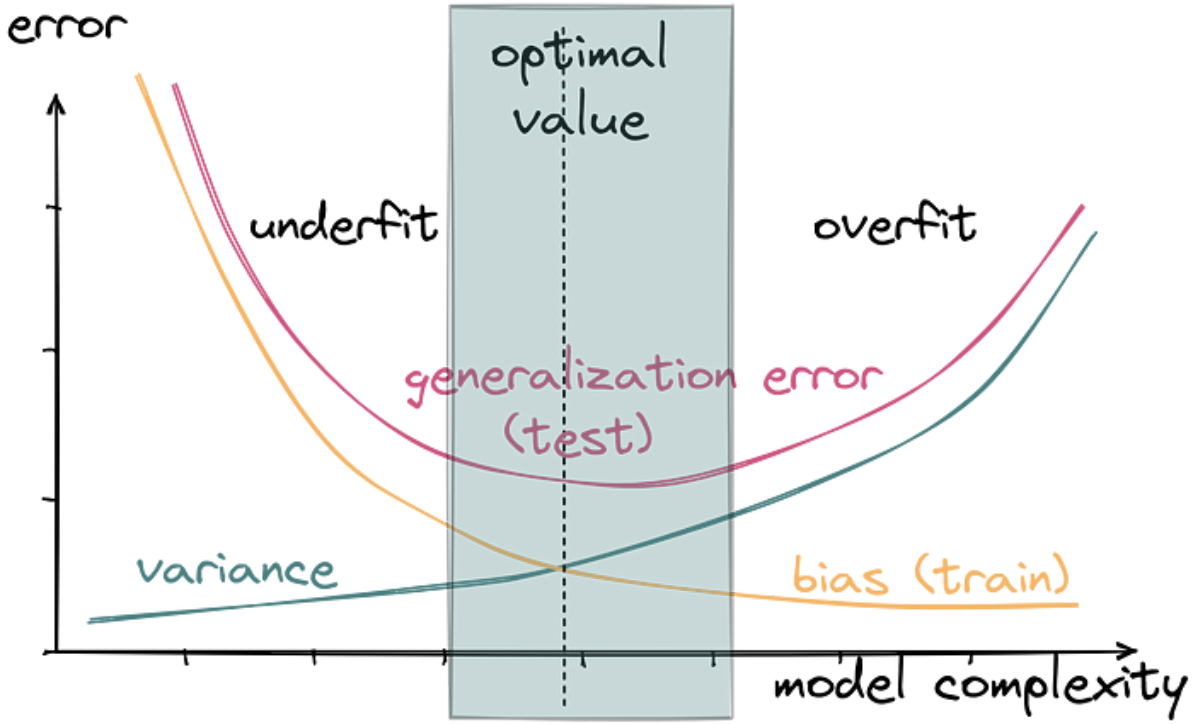
- Most models aim to fit the data the best it can (e.g, OLS)
- The data are the data – the data we have are the best we know of what represents the population

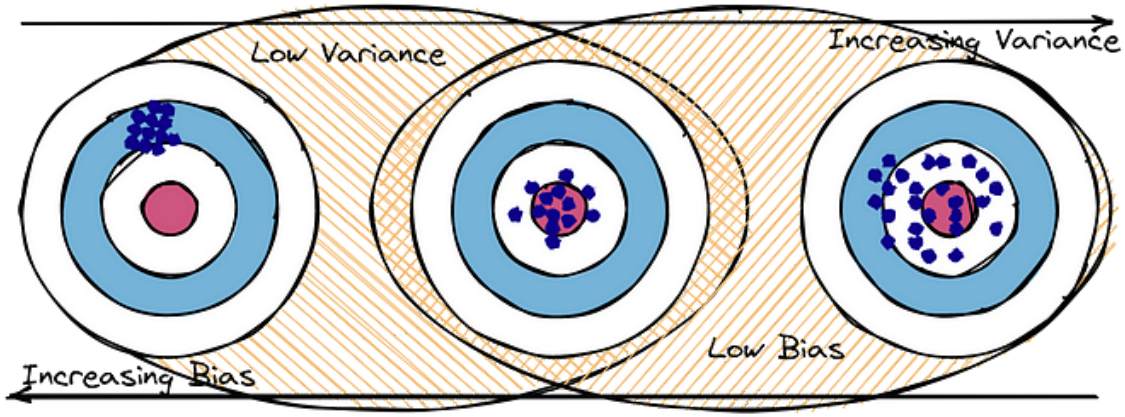
Generalizability

How often does our sample represent the (target) population?

What about our variables representing the appropriate sample space of predictors (e.g., model selection bias)?

Generalizability





Generalizability

Terminology

- **train(ing)**: fitting a model to a dataset
  - Sometimes "updating" or "fine-tuning" existing parameters
- **validation**: using a trained model to predict unseen data
  - Social sciences: ongoing process of testing the model on many other datasets
  - Data science: unseen training data used as part of model "feedback" (e.g., selecting the model that best predicts these data)
- **test(ing)**: using a trained model to predict data held-out completely from training (even newly collected data)

Generalizability

Methods

- data splitting (train/test)
- bootstrap
- k-folds cross-validation

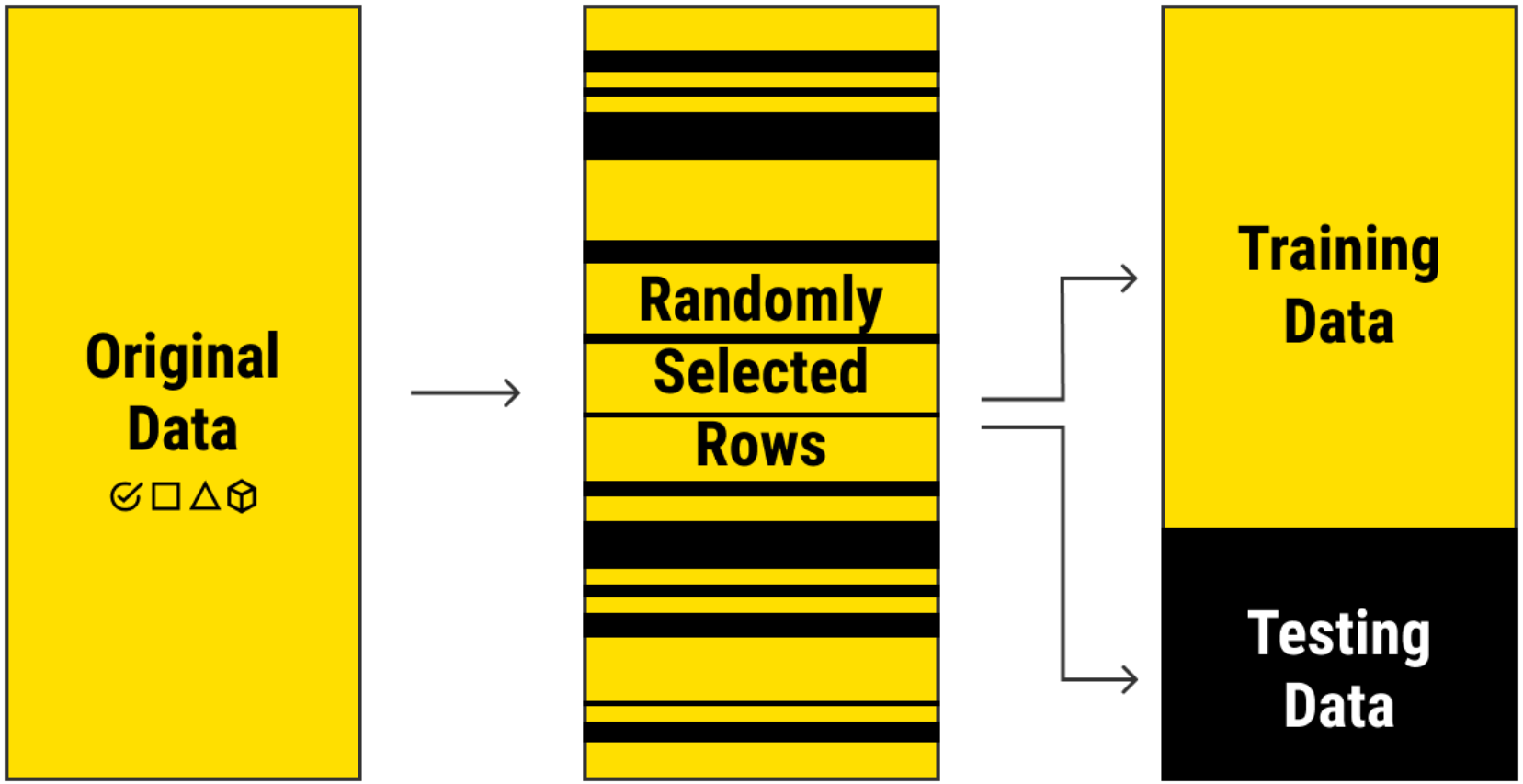
Data Splitting

Data Splitting

Premise

- The data you have are the data you have
- Splitting the data allows you to evaluate generalizability *without* collecting new data
- Provides a single estimate of your model's generalizability

Data Splitting



Data Splitting

- Common splits (train/test): 70/30 and 80/20
- The larger the training dataset, the more likely the model will overfit
- The smaller the training dataset, the more likely the model will underfit

In-class Activity

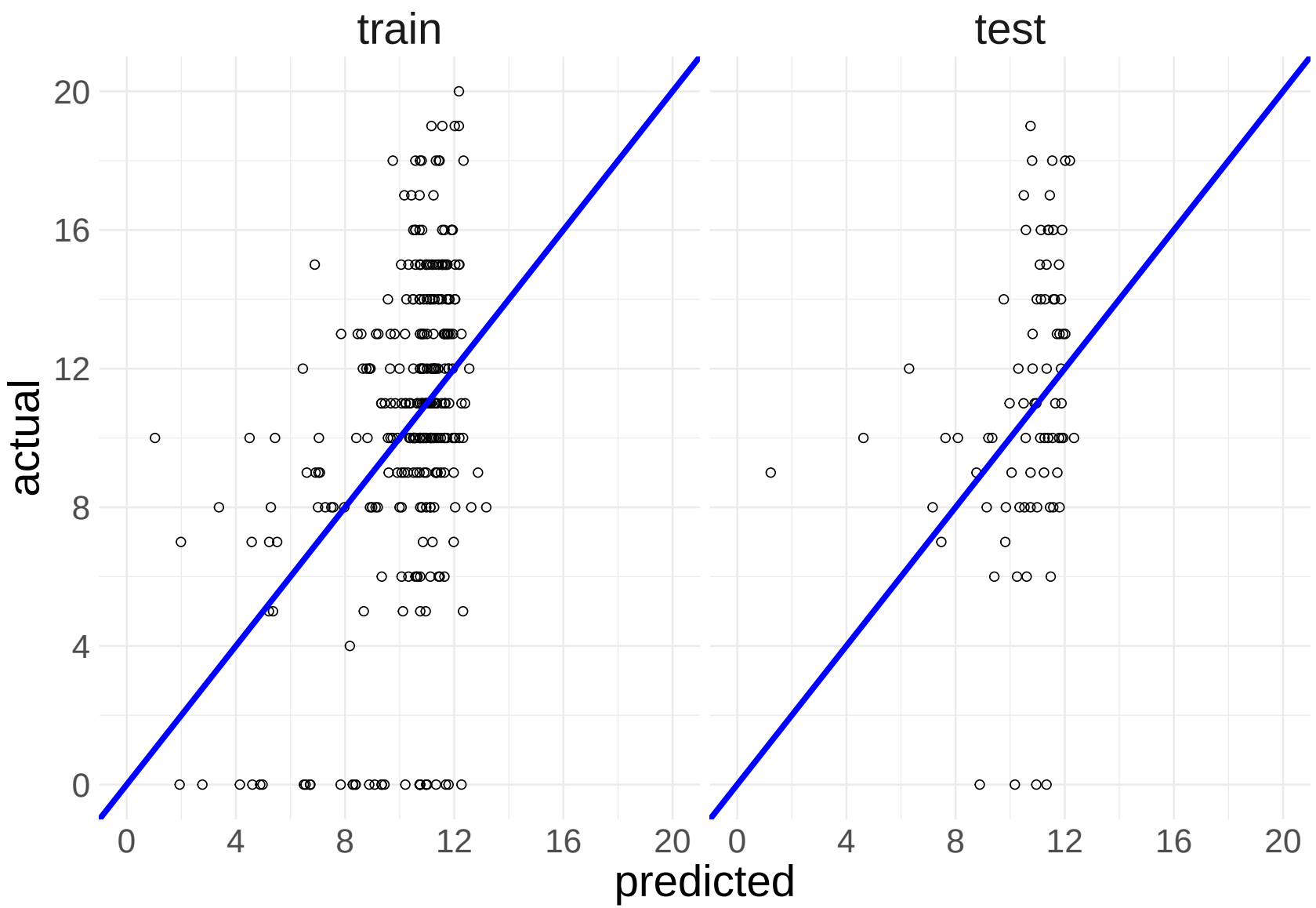
Template in R

```r
# Set seed for reproducibility
set.seed(1234) # don't forget!!

# Get training indices
training <- sample(
  1:nrow(data), # sequence through the cases
  round(nrow(data) * 0.80), # 0.70 or 0.80
  replace = FALSE # do not sample with replacement
)

# Split your data
data_train <- data[training,]
data_test <- data[-training,]

# Fit your model
fit <- model(..., data = data_train)
# replace `model` with your model

# Predict your test data
predict_test <- predict(fit, newdata = data_test)
# Arguments can change depending on function used,
# so RTFM (read the f*cking manual: ?predict.model)

# Perform some evaluation
sqrt(mean((predict_test - data_test$outcome)^2))
# e.g., root mean square error in linear regression
```

Perform Data Splitting

- Use the `student_math_clean.csv` dataset
- Set a seed and split the data 80/20 (train/test)
- Using the **training** dataset, fit a linear model (`lm`) to `final_grade` using the previous variables of interest:
  - `study_time`, `class_failures`, `school_support`, `family_support`, `higher_ed`, `internet_access`, `health`, `absences`
- Using the **testing** dataset, `predict` the values of `final_grade` and compute RMSE

Data Splitting



Data Splitting

Limitations?

Bootstrap

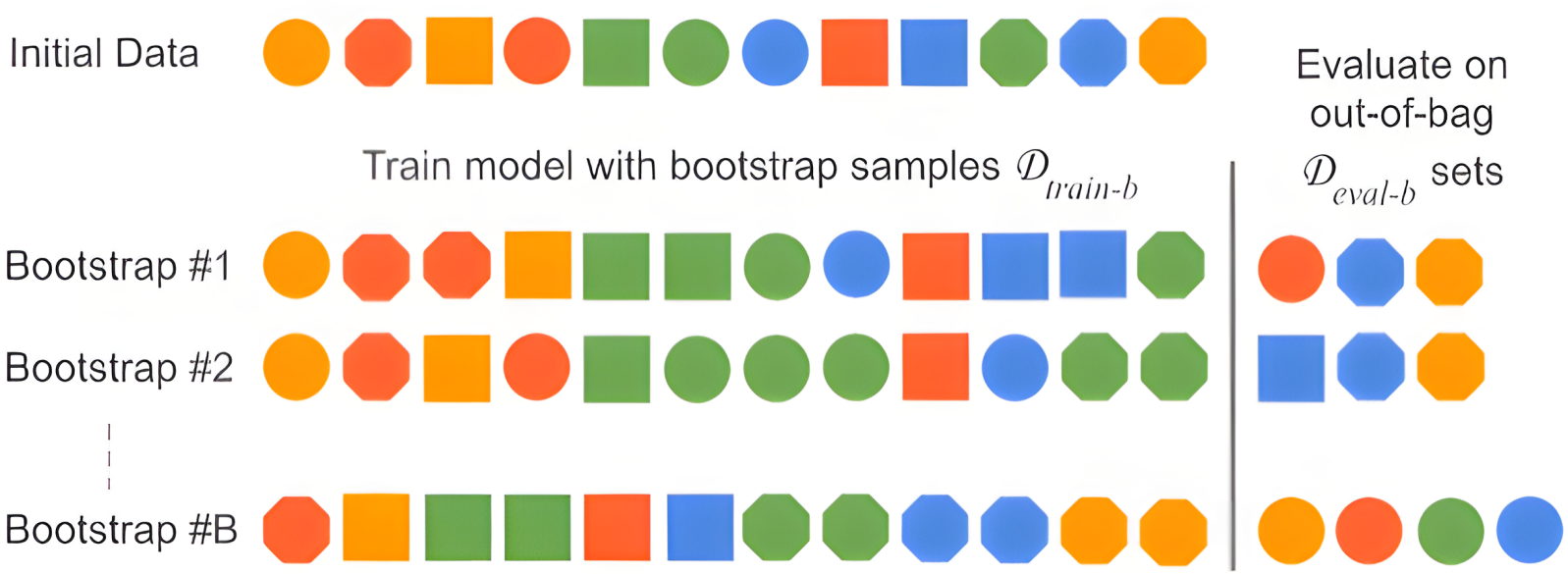
Bootstrap

Premise

- The data you have are the data you have
- Resample from data with replacement with the same N to create a replicate dataset $\tilde{\mathbf{X}}_{train}$
- Cases not included in the replicate dataset are used as the test $\tilde{\mathbf{X}}_{test}$

The bootstrap concept will be relevant for many methods in this course...

Bootstrap


Initial Data
Train model with bootstrap samples $\mathcal{D}_{train\text{-}b}$
Evaluate on out-of-bag $\mathcal{D}_{eval\text{-}b}$ sets
Bootstrap #1
Bootstrap #2
Bootstrap #B

Bootstrap

- Common iterations: 100
- The test dataset $\tilde{\mathbf{X}}_{test}$ will vary in size based on observations (not) included in the train dataset $\tilde{\mathbf{X}}_{train}$
- Reported metric (e.g., RMSE, accuracy) is the mean across bootstrap $\tilde{\mathbf{X}}_{test}$

In-class Activity

Template in R

```r
# Set seed for reproducibility
set.seed(1234) # don't forget!!

# Set up training for {caret}
train_control <- trainControl(
  method = "boot", # training method
  number = 100 # number of bootstraps
)

# Perform bootstrap
data_boot <- train(
  form = ..., # formula input
  data = data_train, # data input
  method = "lm", # linear model
  metric = "RMSE", # default
  trControl = train_control # training parameters
)

# Summary (coefficients will be the same)
data_boot

# Compare to original
sqrt(mean((predict_test - data_test$outcome)^2))
```

Perform Bootstrap

- Use the training dataset
- Set a seed (unless previously set)
- Using the **training** dataset, fit a linear model (`lm`) to `final_grade` using the previous variables of interest:
  - `study_time`, `class_failures`, `school_support`, `family_support`, `higher_ed`, `internet_access`, `health`, `absences`
- Compare the RMSE based on resampling versus data splitting: Which one is lower?

Variability of Generalizability

```
# Descriptive statistics
mean(math_boot$resample$RMSE)
```

```
[1] 4.382653
```

```
median(math_boot$resample$RMSE)
```

```
[1] 4.42525
```

```
sd(math_boot$resample$RMSE)
```

```
[1] 0.2580428
```

```
range(math_boot$resample$RMSE)
```

```
[1] 3.567686 4.866306
```

Bootstrap

Limitations?

Relative Importance

Relative Importance

relative importance: how important each variable is for prediction relative to other variables

- permutates each variable, one-by-one, to determine the proportion of effect shift in metric (differs based on method)
  - Linear model: t-statistic
  - Random forest regression: mean squared error
- particularly useful for interpreting more complicated "black box" algorithms

Relative Importance

```
# Compute relative importance
t(varImp(math_boot)$importance)
```

```
        study_time class_failures school_support family_support
Overall   8.082939            100       13.58962       18.37475
        extra_paid_classes higher_ed internet_access absences
Overall                  0  30.22777       0.4594754 13.29506
```

Relative Importance

Compare with significance

	relative_importance	beta_coefficient	p.value	significance
study_time	8.083	0.229	0.343	n.s.
class_failures	100.000	-2.119	0.000	***
school_support	13.590	-0.979	0.202	n.s.
family_support	18.375	-0.833	0.118	n.s.
extra_paid_classes	0.000	0.246	0.643	n.s.
higher_ed	30.228	2.664	0.024	*
internet_access	0.459	0.329	0.623	n.s.
absences	13.295	0.038	0.208	n.s.

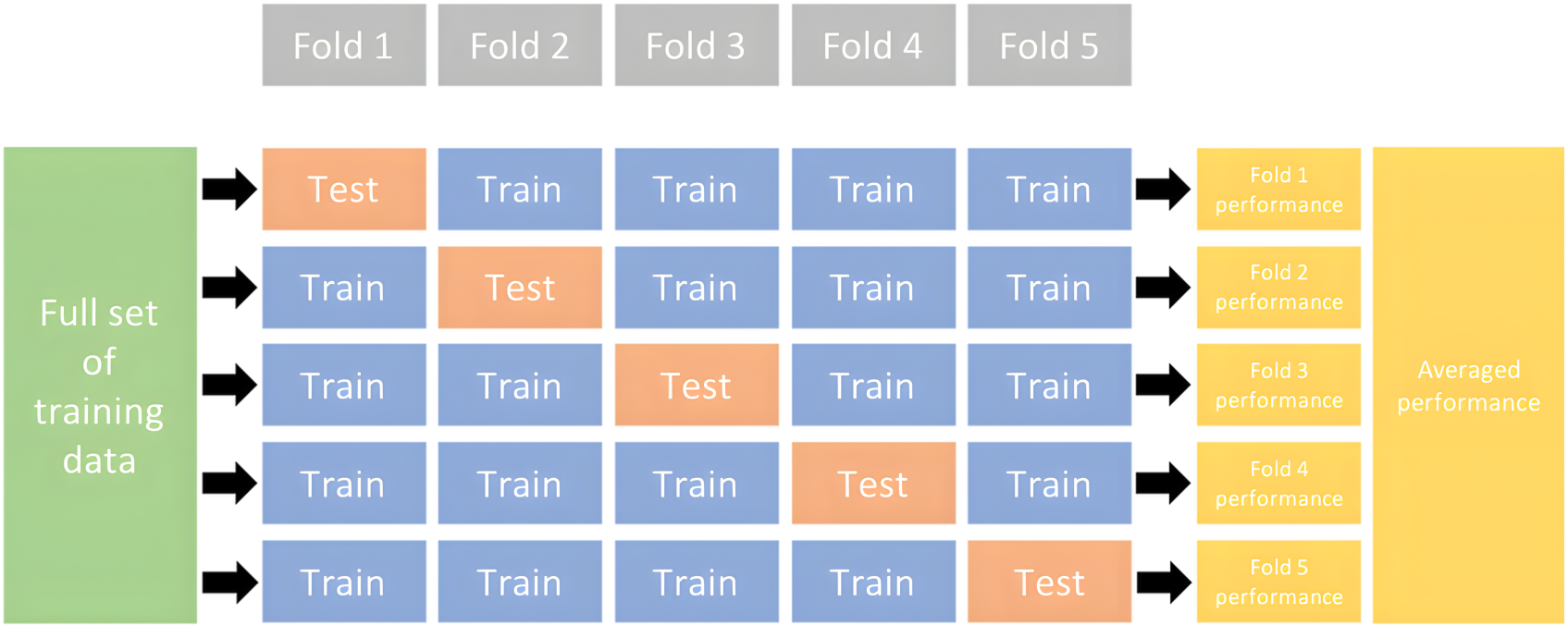
k-folds Cross-validation

k-folds Cross-validation

Premise

- The data you have are the data you have
- Single train/test in data splitting is limited to one split
- Splits can be used in training and testing for the entire dataset

k-folds Cross-validation



k-folds Cross-validation

- Common folds: 5- and 10-folds (equivalent to 80/20 and 90/10 splits)
- Leave-one-out: n-folds (every observation is treated as a test)
- Reported metric (e.g., RMSE, accuracy) is the mean across test splits
- Benefit over other methods: Each observation receives a predicted value (e.g., correlation predicted with actual as a metric)

In-class Activity

Template in R

```r
# Set seed for reproducibility
set.seed(1234) # don't forget!!

# Set up training for {caret}
train_control <- trainControl(
  method = "cv", # training method
  number = 5 # number of folds
)

# Perform bootstrap
data_cv <- train(
  form = ..., # formula input
  data = data_train, # data input
  method = "lm", # linear model
  metric = "RMSE", # default
  trControl = train_control # training parameters
)

# Summary (coefficients will be the same)
data_cv

# Compare to original
sqrt(mean((predict_test - data_test$outcome)^2))
```

Perform 5-fold Cross-validation

- Use the training dataset
- Set a seed (unless previously set)
- Using the **training** dataset, fit a linear model (`lm`) to `final_grade` using the previous variables of interest:
  - `study_time`, `class_failures`, `school_support`, `family_support`, `higher_ed`, `internet_access`, `health`, `absences`
- Compare the RMSE based on k-folds versus data splitting: Which one is lower?

k-folds Cross-validation

Variability of Generalizability

```
# k-folds results
math_cv$resample
```

```
      RMSE   Rsquared      MAE Resample
1 3.965580 0.25593889 3.151722    Fold1
2 4.051995 0.17520479 3.080313    Fold2
3 4.368245 0.18575194 3.421772    Fold3
4 4.917325 0.06231041 3.679247    Fold4
5 4.366784 0.09779386 3.216993    Fold5
```

```
Mean:  4.334
```

```
Median:  4.367
```

```
SD:  0.373
```

Compare with bootstrap and significance

	cv_rellmp	boot_rellmp	beta	p.value	stars
study_time	8.083	8.083	0.229	0.343	n.s.
class_failures	100.000	100.000	-2.119	0.000	***
school_support	13.590	13.590	-0.979	0.202	n.s.
family_support	18.375	18.375	-0.833	0.118	n.s.
extra_paid_classes	0.000	0.000	0.246	0.643	n.s.
higher_ed	30.228	30.228	2.664	0.024	*
internet_access	0.459	0.459	0.329	0.623	n.s.
absences	13.295	13.295	0.038	0.208	n.s.

k-folds Cross-validation

Limitations?

At Home Activity

Following the in-class activities, perform:

- Data splitting
- Bootstrap
- k-folds cross-validation

using **logistic** regression to predict `extra_paid_classes`

Readings for Next Week

Readings

- ESL Chapters: 3.4, 3.4.1, 3.4.2, and 4.4.4
- HML: Chapter 6

Optional

- Jacobucci et al. - 2016
- Seeboth and Mõttus - 2018