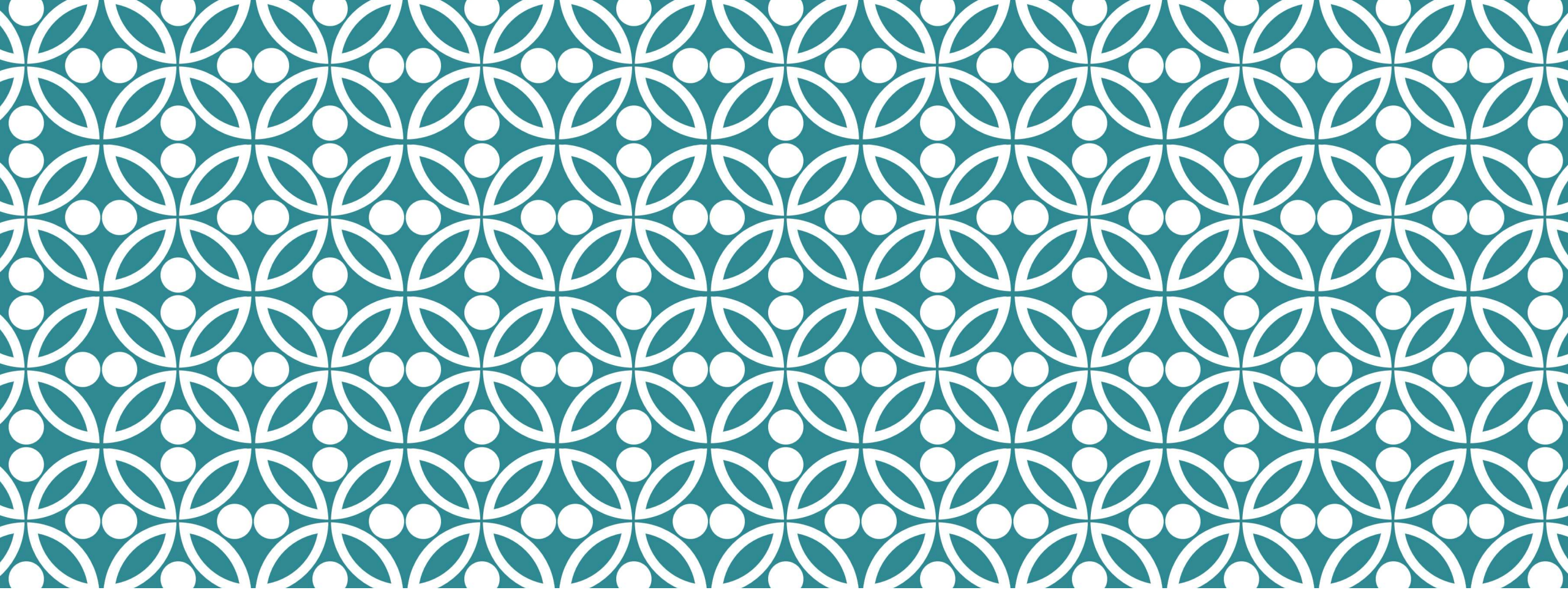


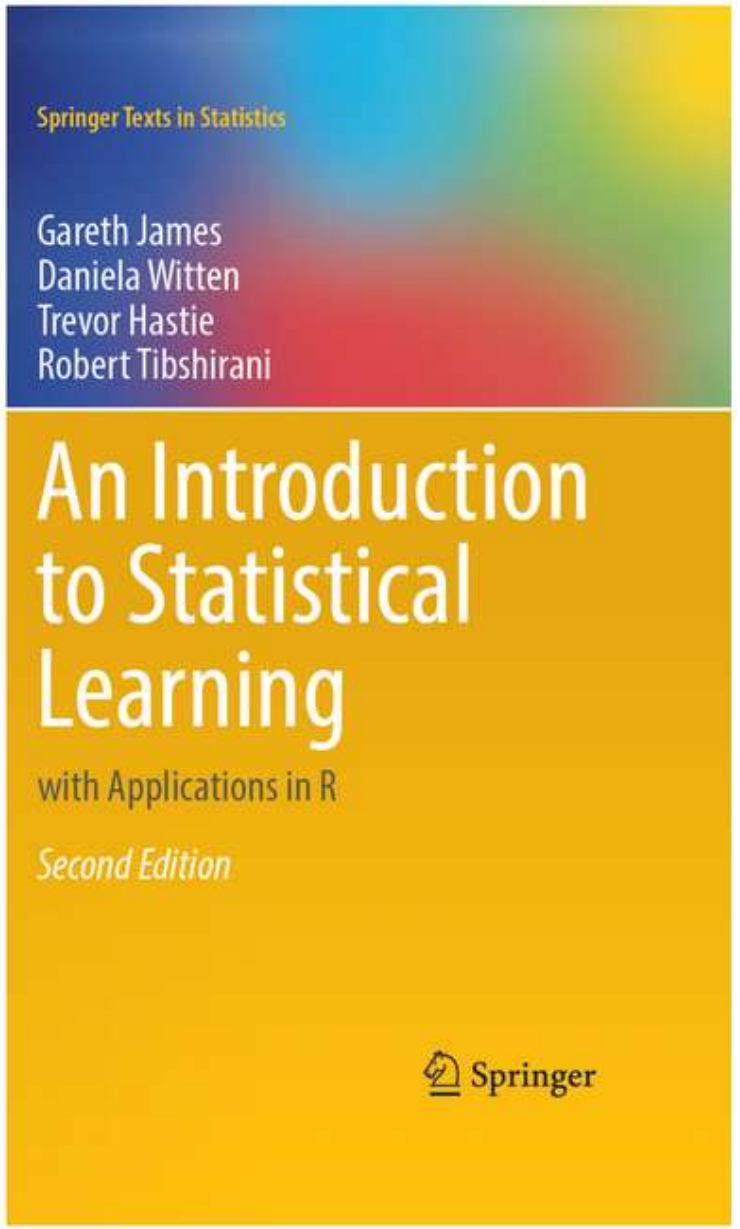
SESSION 5.2: SUPERVISED LEARNING

DR. SOFIA GIL-CLAVEL

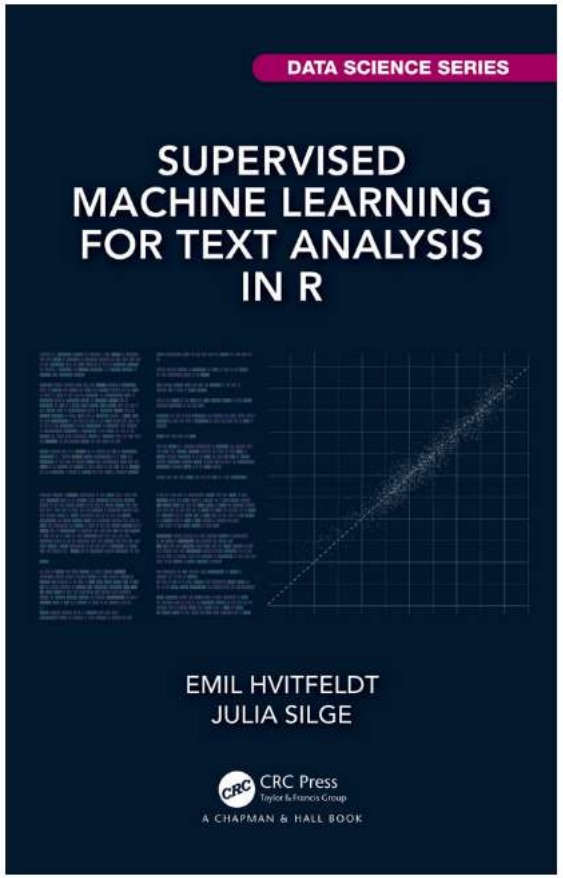
- Recap: unsupervised learning
- Supervised learning
- Assessing models and results
- Text classification

WE WILL BE FOLLOWING:

You can find the book for free here:
https://www.statlearning.com/



AND . . . HTTPS://SMLTAR.COM/



Modeling as a statistical practice can encompass a wide variety of activities. This book focuses on *supervised or predictive modeling for text*, using text data to make predictions about the world around us. We use the tidymodels framework for modeling, a consistent and flexible collection of R packages developed to encourage good statistical practice.

Supervised machine learning using text data involves building a statistical model to estimate some output from input that includes language. The two types of models we train in this book are regression and classification. Think of regression models as predicting numeric or continuous outputs, such as predicting the year of a United States Supreme Court opinion from the text of that opinion. Think of classification models as predicting outputs that are discrete quantities or class labels, such as predicting whether a GitHub issue is about documentation or not from the text of the issue. Models like these can be used to make predictions for new observations, to understand what features or characteristics contribute to differences in the output, and more. We can evaluate our models using performance metrics to determine which are best, which are acceptable for our specific context, and even which are fair.

1. RECAP: UNSUPERVISED LEARNING

UNSUPERVISED VS. SUPERVISED

$$? \sim X1_i + X2_i + X3_i$$

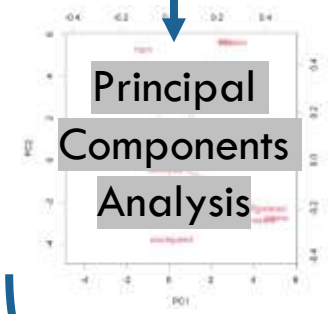
Unsupervised

$$\not{Y}_i \sim X1_i + X2_i + X3_i$$

Numeric Matrix

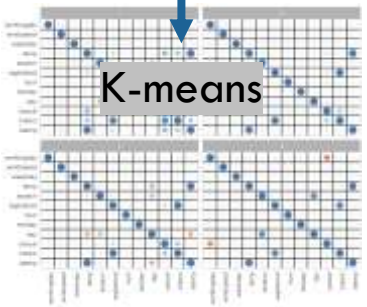
Dimensionality reduction

Principal Components Analysis



Clustering

K-means

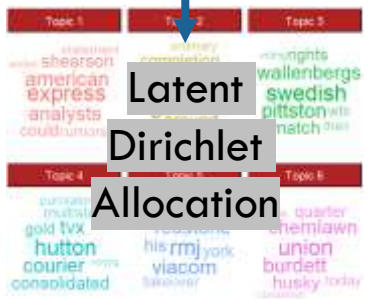


Text

Document-Term Matrix

Topic Modelling

Latent Dirichlet Allocation



Supervised

$$Y_i \sim X1_i + X2_i + X3_i$$

Part of exploratory data analysis.

2. SUPERVISED LEARNING

1. Regression vs. Classification
2. Prediction vs. Inference
3. Some classification models

SOME NOTATION

$$Y_i \sim X1_i + X2_i + X3_i \sim f(X_i)$$

```
> Bikeshare
# A tibble: 8,645 x 12
   bikers season   day holiday weekday workingday  temp atemp   hum
   <dbl>  <dbl> <dbl>   <dbl>   <dbl>      <dbl> <dbl> <dbl> <dbl>
 1     16      1     1       0       6          0  0.24 0.288  0.81
 2     40      1     1       0       6          0  0.22 0.273  0.8
 3     32      1     1       0       6          0  0.22 0.273  0.8
 4     13      1     1       0       6          0  0.24 0.288  0.75
 5      1      1     1       0       6          0  0.24 0.288  0.75
 6      1      1     1       0       6          0  0.24 0.258  0.75
 7      2      1     1       0       6          0  0.22 0.273  0.8
 8      3      1     1       0       6          0  0.2  0.258  0.86
 9      8      1     1       0       6          0  0.24 0.288  0.75
10     14      1     1       0       6          0  0.32 0.348  0.76
# i 8,635 more rows
# i 3 more variables: windspeed <dbl>, casual <dbl>,
#   registered <dbl>
# i Use `print(n = ...)` to see more rows
```

dependent variable (Y)

independent variables (X)

This value comes from our data.

We estimate the model (f) parameters $\widehat{\beta_i}$ and use them to predict $\widehat{Y}_i$

$$\widehat{\beta_1} X1_i + \widehat{\beta_2} X2_i + \widehat{\beta_3} X3_i = \widehat{Y}_i$$

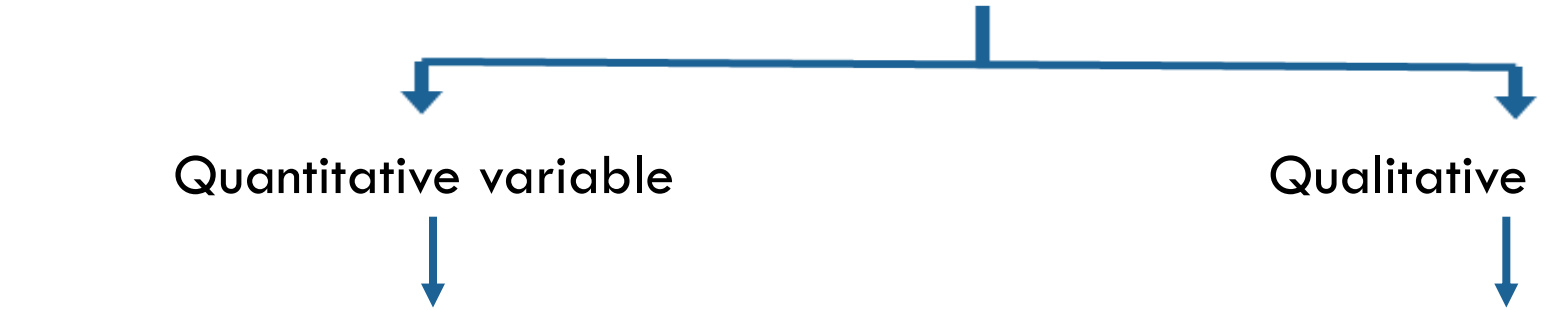
This value is the prediction.

2.1 REGRESSION VS. CLASSIFICATION

REGRESSION VS. CLASSIFICATION

Y_i

Take a moment to think about the type of values Y can get.



Quantitative variable

- Discrete:
  - Person's age
  - Day of the year
  - Number of cats per neighborhood
- Continuous
  - Height
  - Income
  - A house value

Qualitative variable

- Dichotomous:
  - Person's marital status (married or not)
  - Gender (female or not)
- Ordinal:
  - 5 stars rank
- Categorical:
  - Country name
  - Brand of a product

REGRESSION VS. CLASSIFICATION

Y_i

Take a moment to think about the type of values Y can get.

Quantitative variable

↓

Discrete
Continuous

↓ Then the model is

$f(X_i)$

Regression

↓ e.g.

Least squares linear regression

Qualitative variable

↓

Dichotomous
Ordinal
Categorical

↓ Then the model is

$f(X_i)$

Classification

↓ e.g.

Logistic regression

REGRESSION VS. CLASSIFICATION

Y_i

Take a moment to think about the type of values Y can get.

Quantitative variable

↓

Discrete
Continuous

↓ Then the model is

$f(X_i)$

Regression

↓ e.g.

Least squares linear regression

Qualitative variable

↓

Dichotomous

↓ Then the model is

$f(X_i)$

Classification

↓ e.g.

Logistic regression

2.2 PREDICTION VS. INFERENCE

PREDICTION VS. INFERENCE

$$Y_i \sim X1_i + X2_i + X3_i$$

$$\downarrow f(X_i)$$

$$\hat{Y}_i = \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \hat{\beta}_3 X3_i$$

Prediction

Predict likelihood someone gets and allergic reaction to a medicine.

The focus is on the predicted value:

$\hat{Y}_i$

We care about the **accuracy** of the model.

Inference

Understand the **relationship between** income and the variables years of education and seniority.

The focus is on the coefficients:

$\hat{\beta}_j$

We care about the **interpretability** of the model.

PREDICTION VS. INFERENCE

$$Y_i \sim X1_i + X2_i + X3_i$$

$$\downarrow f(X_i)$$

$$\hat{Y}_i = \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \hat{\beta}_3 X3_i$$

Prediction

Predict likelihood someone gets and allergic reaction to a medicine.

↓

The focus is on the predicted value:

$\hat{Y}_i$

↓

We care about the **accuracy** of the model.

Inference

Understand the **relationship between** income and the variables years of education and seniority.

↓

The focus is on the coefficients:

$\hat{\beta}_j$

↓

We care about the **interpretability** of the model.

IN SUMMARY

Y_i → $f(X_i)$ Classification → **Prediction** → $\sum \hat{\beta}_j X j_i = \hat{Y}_i$

Qualitative variable

↓

Dichotomous

↓ Then the model is

$f(X_i)$

Classification

↓ e.g.

Logistic regression

Predict likelihood someone gets and allergic reaction to a medicine.

↓

The focus is on the predicted value:

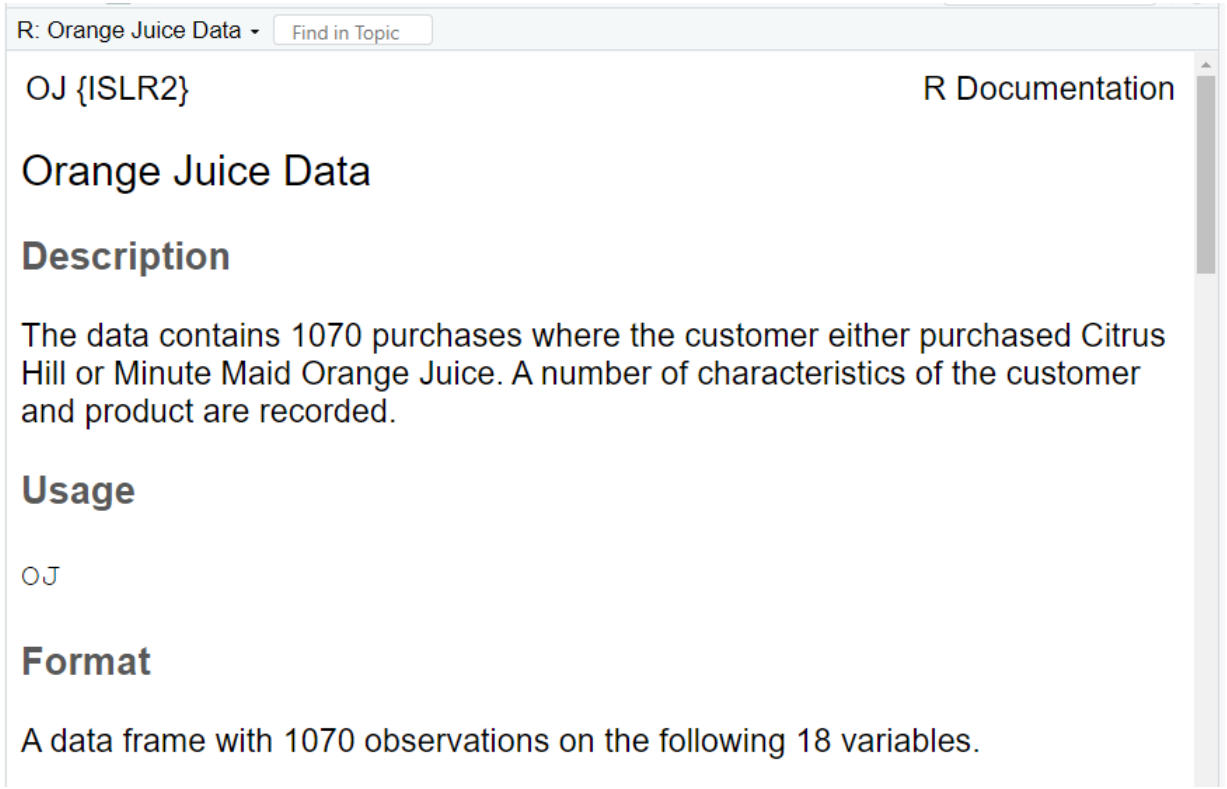
$\hat{Y}_i$

↓

We care about the **accuracy** of the model.

2.3 SOME CLASSIFICATION MODELS

THE DATA



R: Orange Juice Data Find in Topic

OJ {ISLR2} R Documentation

Orange Juice Data

Description

The data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of the customer and product are recorded.

Usage

```
OJ
```

Format

A data frame with 1070 observations on the following 18 variables.

- **Purchase:** A factor with levels CH and MM indicating whether the customer purchased Citrus Hill or Minute Maid Orange Juice
- **WeekofPurchase:** Week of purchase
- **StoreID:** Store ID
- **PriceCH:** Price charged for CH
- **PriceMM:** Price charged for MM
- **DiscCH:** Discount offered for CH
- **DiscMM:** Discount offered for MM
- **SpecialCH:** Indicator of special on CH
- **SpecialMM:** Indicator of special on MM
- **LoyalCH:** Customer brand loyalty for CH
- **SalePriceMM:** Sale price for MM
- **SalePriceCH:** Sale price for CH
- **PriceDiff:** Sale price of MM less sale price of CH
- **Store7:** A factor with levels No and Yes indicating whether the sale is at Store 7
- **PctDiscMM:** Percentage discount for MM
- **PctDiscCH:** Percentage discount for CH
- **ListPriceDiff:** List price of MM less list price of CH
- **STORE:** Which of 5 possible stores the sale occurred at

CLASSIFICATION

Classifying means predicting a qualitative response for an observation, since it involves assigning the observation to a category, or class.

There are many possible classification techniques, or classifiers, that one might use to predict a qualitative response.

We will focus on some of the widely-used classifiers:

- Logistic regression
- Linear discriminant analysis

LOGISTIC REGRESSION

This model predicts the probability of a category given the independent variables:

$$P(Y = MM|X) = p(X)$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \quad \rightarrow \quad \hat{Y}_i = \begin{cases} MM \text{ if } p(X) > 0.5 \\ CH \text{ if } p(X) \leq 0.5 \end{cases}$$

LINEAR DISCRIMINANT ANALYSIS

Logistic regression involves directly modeling $P(Y = MM|X)$. In statistical jargon, we model the conditional distribution of the response Y, given the predictor(s) X.

In LDA, we model the distribution of the predictors X separately in each of the response classes (i.e. for each value of Y).

We then use Bayes' theorem to flip these around into estimates for $P(Y = MM|X = x)$.

When the distribution of X within each class is assumed to be normal, it turns out that the model is very similar in form to logistic regression.

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}.$$

5 MINS BREAK

3. ASSESSING THE RESULTS

1. Assessing model accuracy
2. Model evaluation: resampling using cross validation
3. The plan

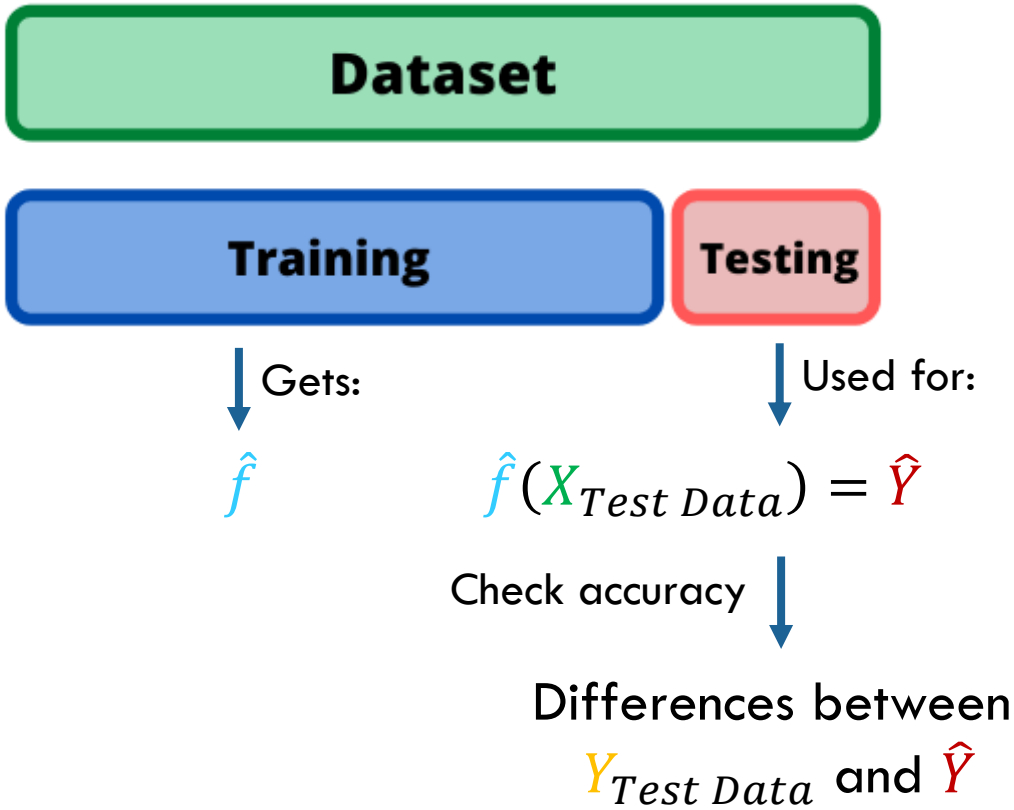
3.1 ASSESSING MODEL ACCURACY

TRAIN AND TEST DATA

The training data is the data observations to train, or teach, our method how to estimate $\hat{f}$.

Our goal is to apply a statistical learning method to the training data to estimate the unknown function $\hat{f}$. In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X_i)$.

But in general, we do not really care how well the function $\hat{f}$ works on the training data. Rather, we are interested in the **accuracy of the predictions** ($\hat{Y}$) that we **obtain when we apply our trained model to** previously unseen data, i.e. **the test data**.



ASSESSING MODEL ACCURACY

The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the ***test error rate***, the proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the test observations:

$$Ave\left(I\left(Y_i \neq \hat{Y}_i\right)\right)$$

A good classifier is one for which the test error is smallest.

3.2 MODEL EVALUATION: RESAMPLING USING CROSS VALIDATION

MODEL EVALUATION

What if we want to compare models or different model parameters. If we use the test set for these kinds of tasks, we risk fooling ourselves that we are doing better than we really are.

Another option for evaluating models is to **predict one time on the training set to measure performance**. This is the *same data* that was used to train the model. Evaluating on the training data often results in performance estimates that are too optimistic. This is especially true for powerful machine learning algorithms that can learn subtle patterns from data; we risk overfitting to the training set. Then what?

Source: https://smltar.com/mlclassification#mlclassification

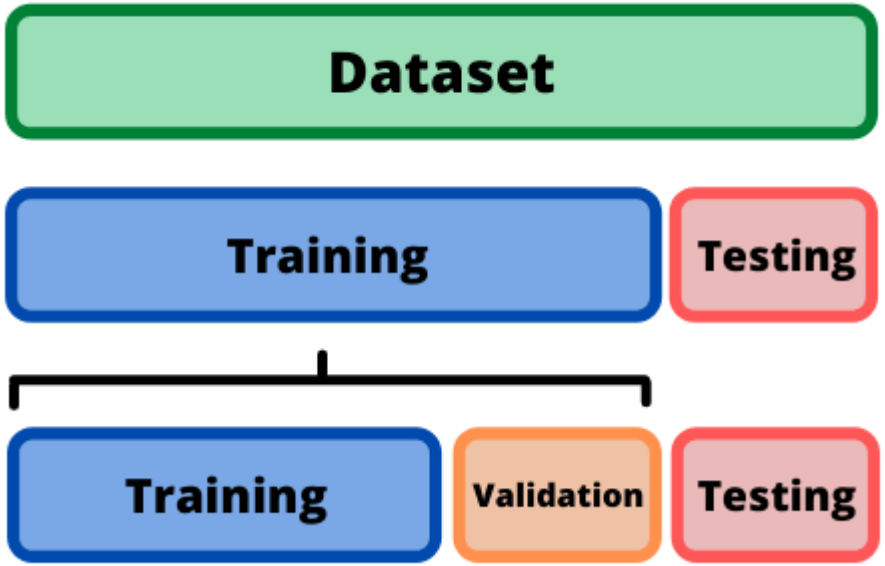
MODEL EVALUATION

Yet another option for evaluating or comparing models is to use a separate **validation set.**

Source: https://smltar.com/mlclassification#mlclassification

MODEL EVALUATION

Yet another option for evaluating or comparing models is to use a separate **validation set**. In this situation, we split our data *not* into two sets (training and testing) but into three sets (testing, training, and validation). **The validation set is used for computing performance metrics to compare models or model parameters.**



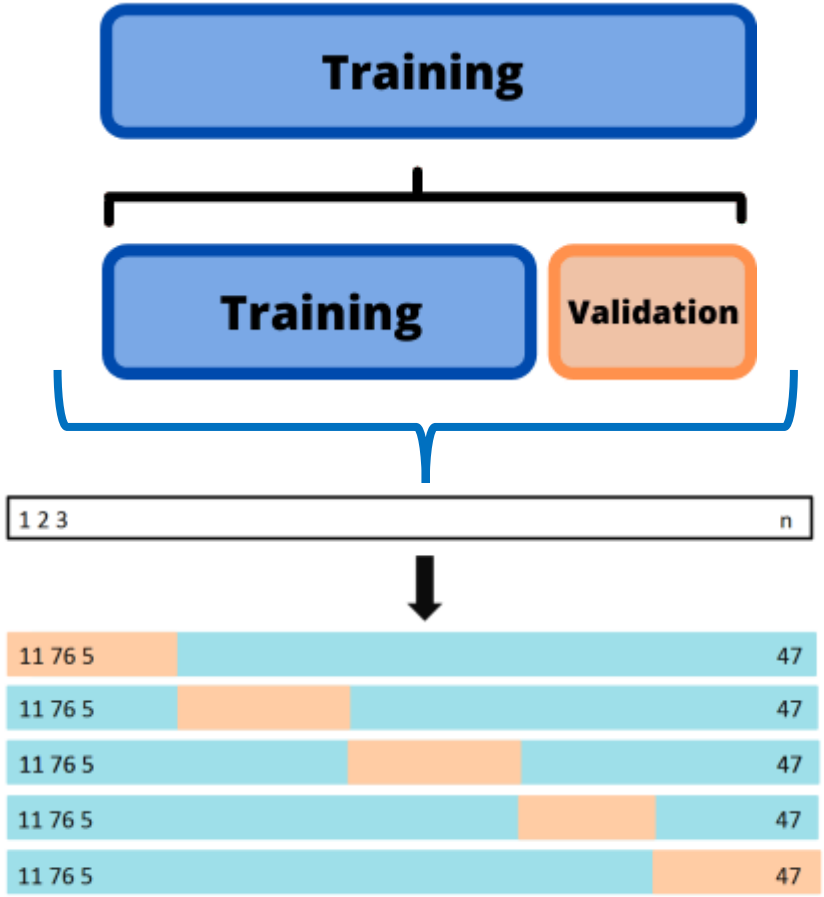
Source: https://smltar.com/mlclassification#mlclassification

RESAMPLING

What are we to do, then, if we want to train multiple models and find the best one? Or compute a reliable estimate for how our model has performed without wasting the valuable testing set?

We can use **resampling**: holding out a subset of the training observations from the fitting process and then applying the statistical learning method to those held out observations.

Resampling is also useful to estimate accuracy when you do not have a big enough dataset to split into training and testing.



K-FOLD CROSS-VALIDATION

This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size.

1. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.
2. The accuracy is computed and saved, MSE1, on the observations in the held-out fold.
3. This procedure is repeated k times; each time, a different group of observations is treated as a validation set.

This process results in k estimates of the test error, MSE1, MSE2, . . . , MSEk. The k-fold CV estimate is computed by averaging these values:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i.$$

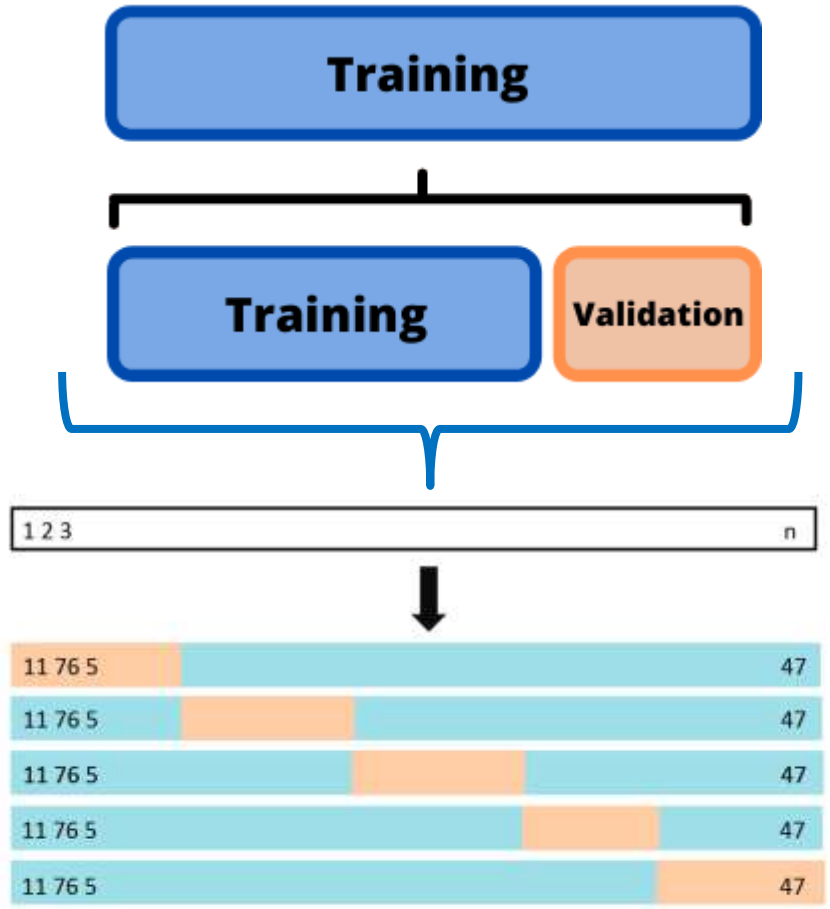


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

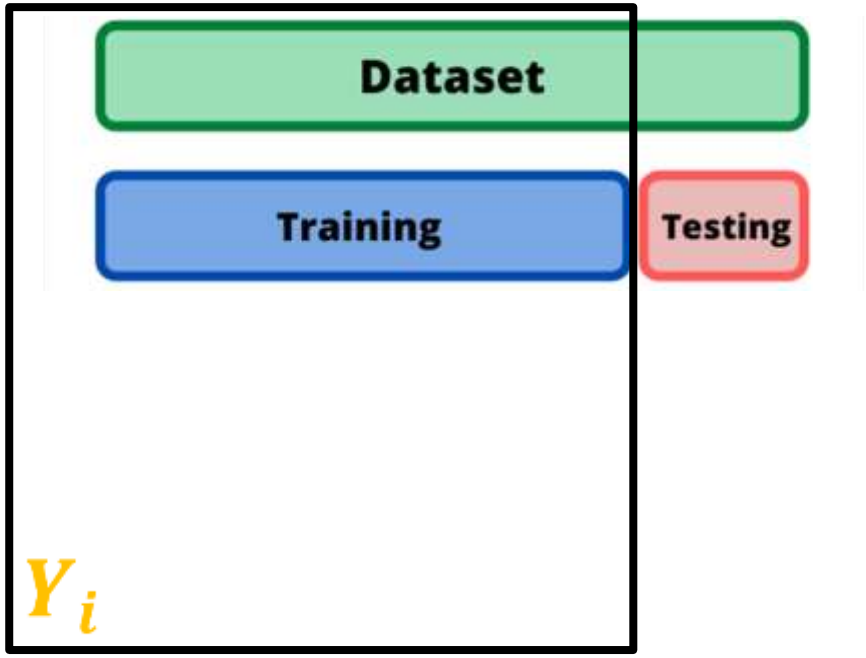
Source: https://www.statlearning.com/

3.3 THE PLAN

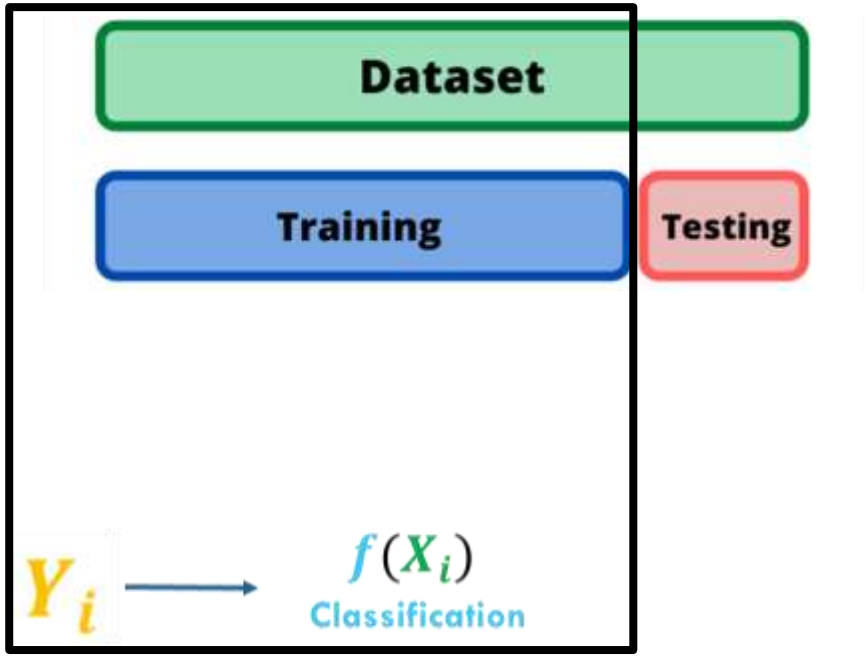
THE PLAN

THE PLAN

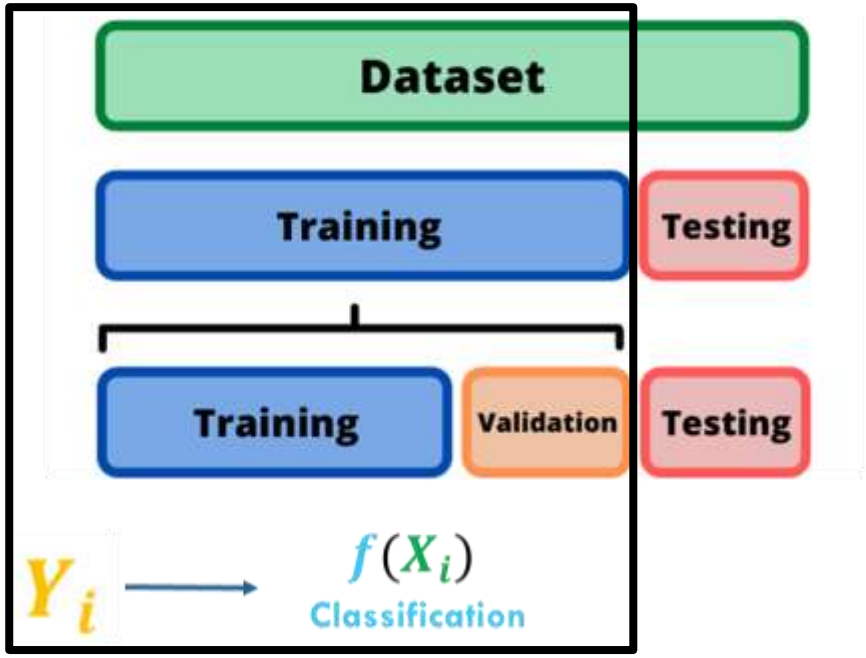
THE PLAN



THE PLAN



THE PLAN


Dataset
Training
Testing
Training
Validation
Testing
Y_i
$f(X_i)$
Classification

THE PLAN

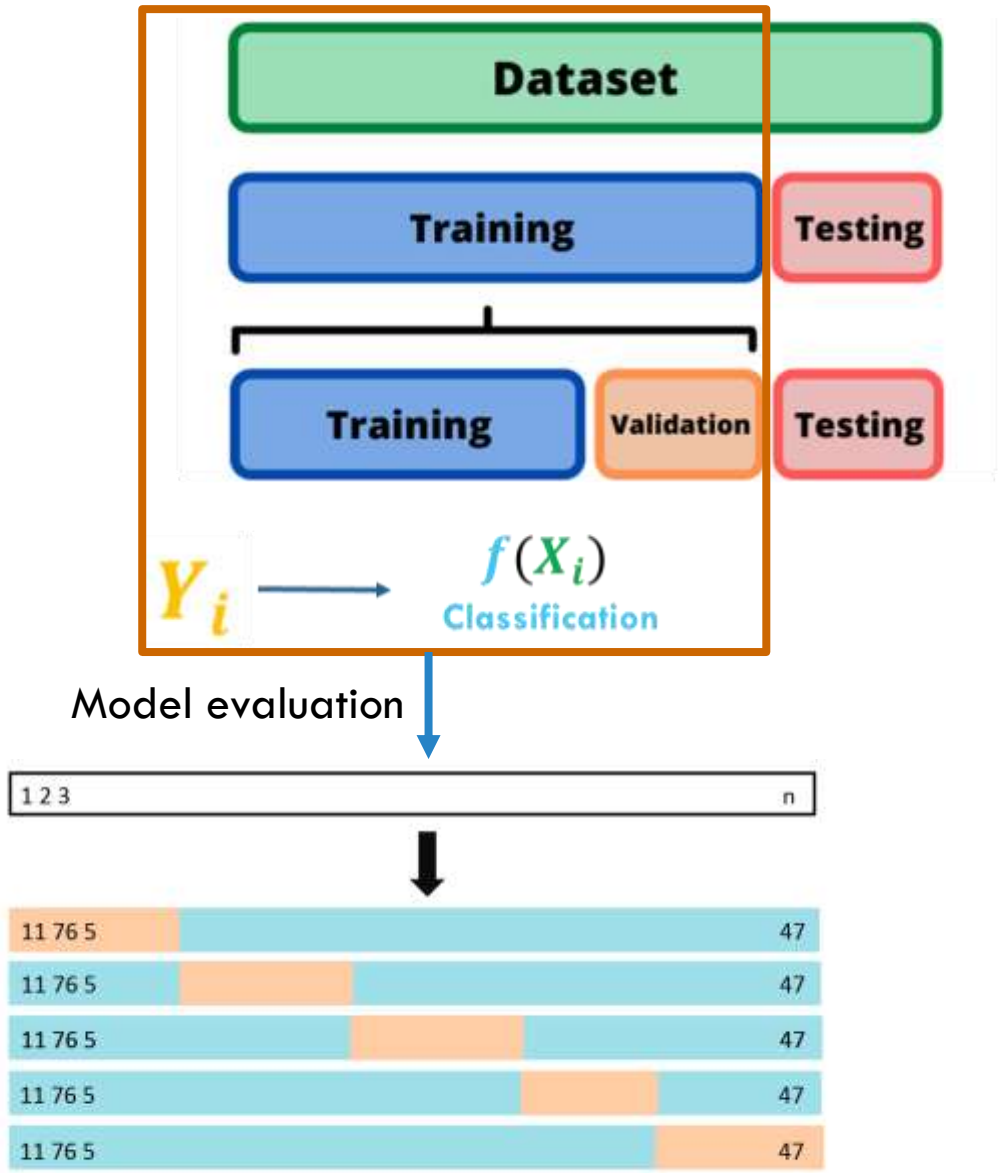
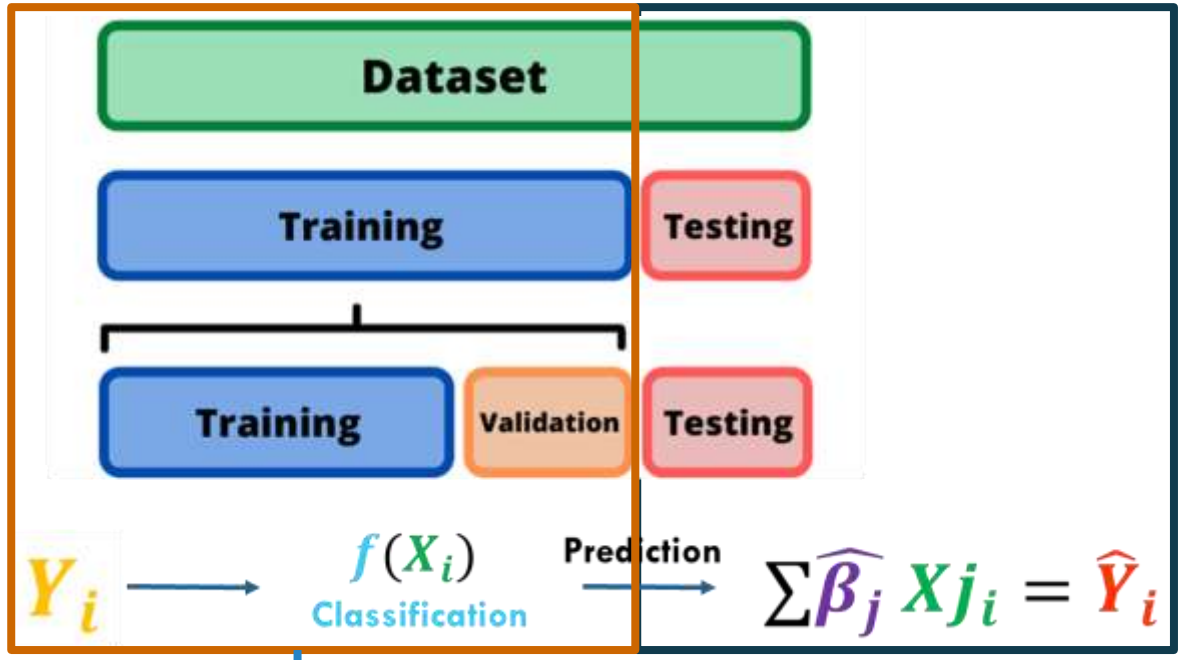


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

THE PLAN



Model evaluation

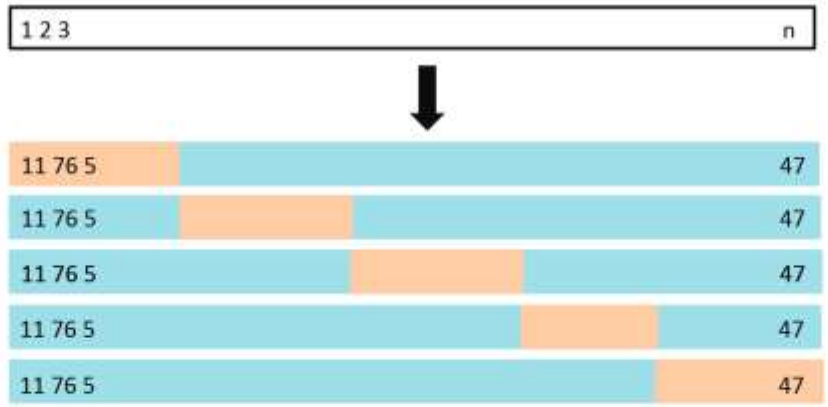
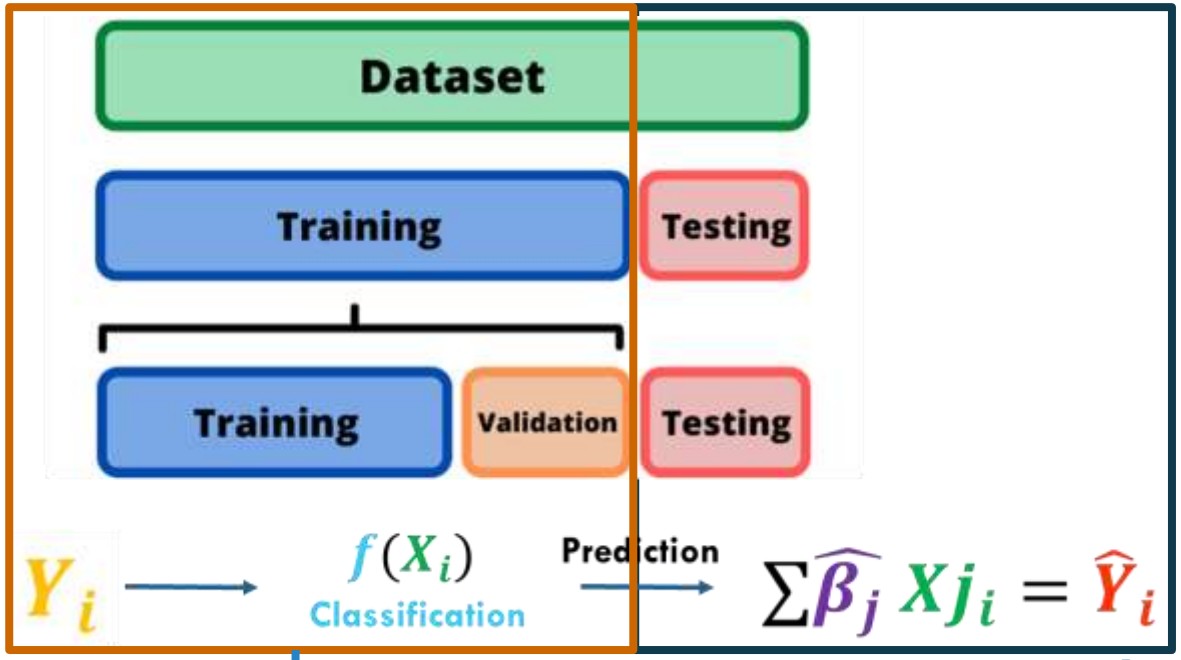
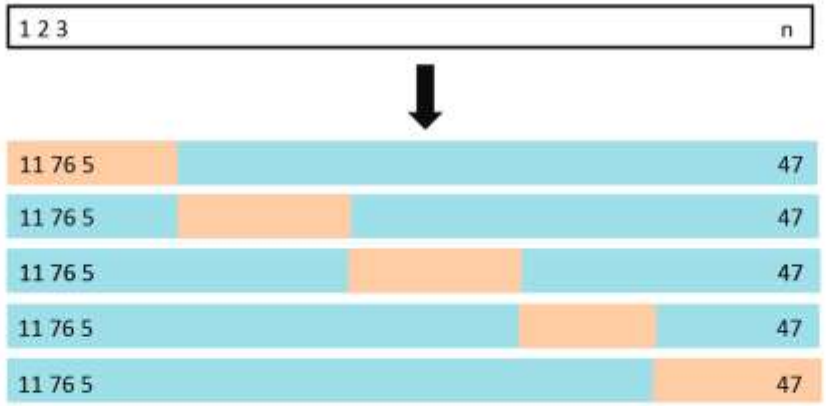


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

THE PLAN



Model evaluation



Model accuracy on the final model

$$Ave\left(I\left(Y_i \neq \hat{Y}_i\right)\right)$$

FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*


OK! WE HAVE THE PLAN!!
LET'S DO THIS!
imgflip.com
moviefone

5 MINS BREAK

4. TEXT CLASSIFICATION

THE DATA

```
> complaints
# A tibble: 117,214 × 18
   date_received product             sub_product issue sub_issue consumer_complaint_n…¹
   <date>        <chr>               <chr>       <chr> <chr>     <chr>
 1 2019-09-24    Debt collection     I do not k… Atte… Debt is … "transworld systems i…
 2 2019-10-25    Credit reporting, … Credit rep… Inco… Informat… "I would like to requ…
 3 2019-11-08    Debt collection     I do not k… Comm… Frequent… "Over the past 2 week…
 4 2019-09-08    Money transfer, vi… Domestic (… Frau… NA        "I was sold access to…
 5 2019-09-24    Debt collection     Medical de… Atte… Debt is … "While checking my cr…
 6 2019-11-20    Credit reporting, … Credit rep… Inco… Account … "I would like the cre…
 7 2019-09-19    Credit reporting, … Credit rep… Inco… Personal… "MY NAME IS XXXX XXXX…
 8 2019-11-22    Credit reporting, … Credit rep… Inco… Personal… "Today XX/XX/XXXX wen…
 9 2019-04-17    Credit reporting, … Credit rep… Prob… Their in… "XXXX is reporting in…
10 2019-10-24    Credit reporting, … Credit rep… Prob… Their in… "Please reverse the l…
# ℹ 117,204 more rows
# ℹ abbreviated name: ¹consumer_complaint_narrative
# ℹ 12 more variables: company_public_response <chr>, company <chr>, state <chr>,
#   zip_code <chr>, tags <chr>, consumer_consent_provided <chr>, submitted_via <chr>,
#   date_sent_to_company <date>, company_response_to_consumer <chr>,
#   timely_response <chr>, consumer_disputed <chr>, complaint_id <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

Download from here:
https://github.com/EmilHvitfeldt/smltar/blob/master/data/complaints.csv.gz

THE PIPELINE

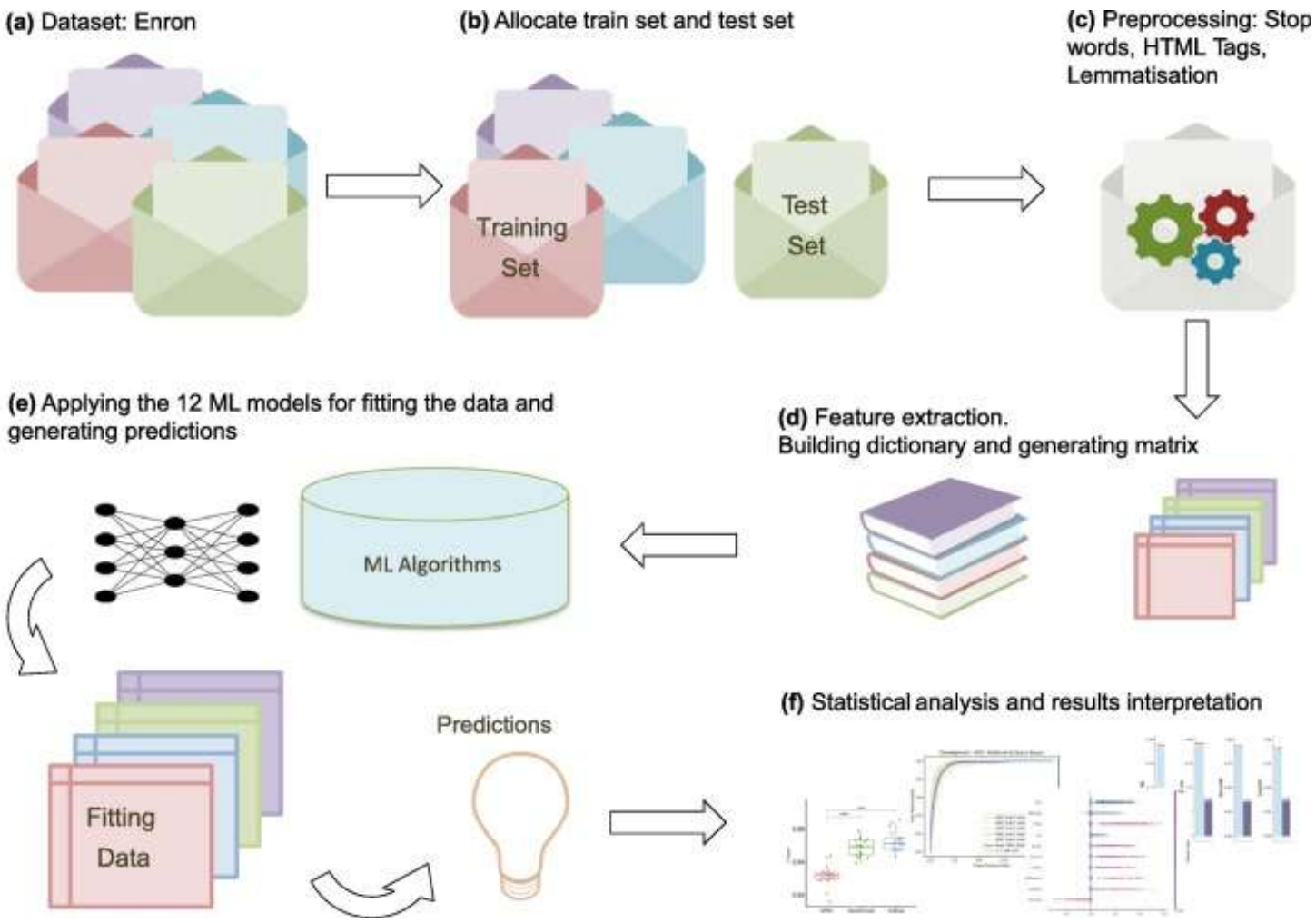


Fig. 1. **Pipeline diagram of the preprocessing and classification steps.** The pipeline shows the process applied for investigating the performance of 12 classifiers. (a) The Enron dataset is selected for the study; (b) the train and test sets are allocated (70% train set and 30% test set); (c) the preprocessing stage is then applied, including removal of stop words, HTML tag and lemmatisation; (d) the features are then extracted for generating a dictionary, based on the number of most occurring features; (e) finally, the 12 algorithms receive the matrices for fitting the data and predicting the classification outcomes; (f) comparative analysis is then performed to assess the statistical significance of the results and provide an accurate interpretation of the machine learning classification outcomes.

Source: https://www.sciencedirect.com/science/article/pii/S0957417422005802

THE PIPELINE

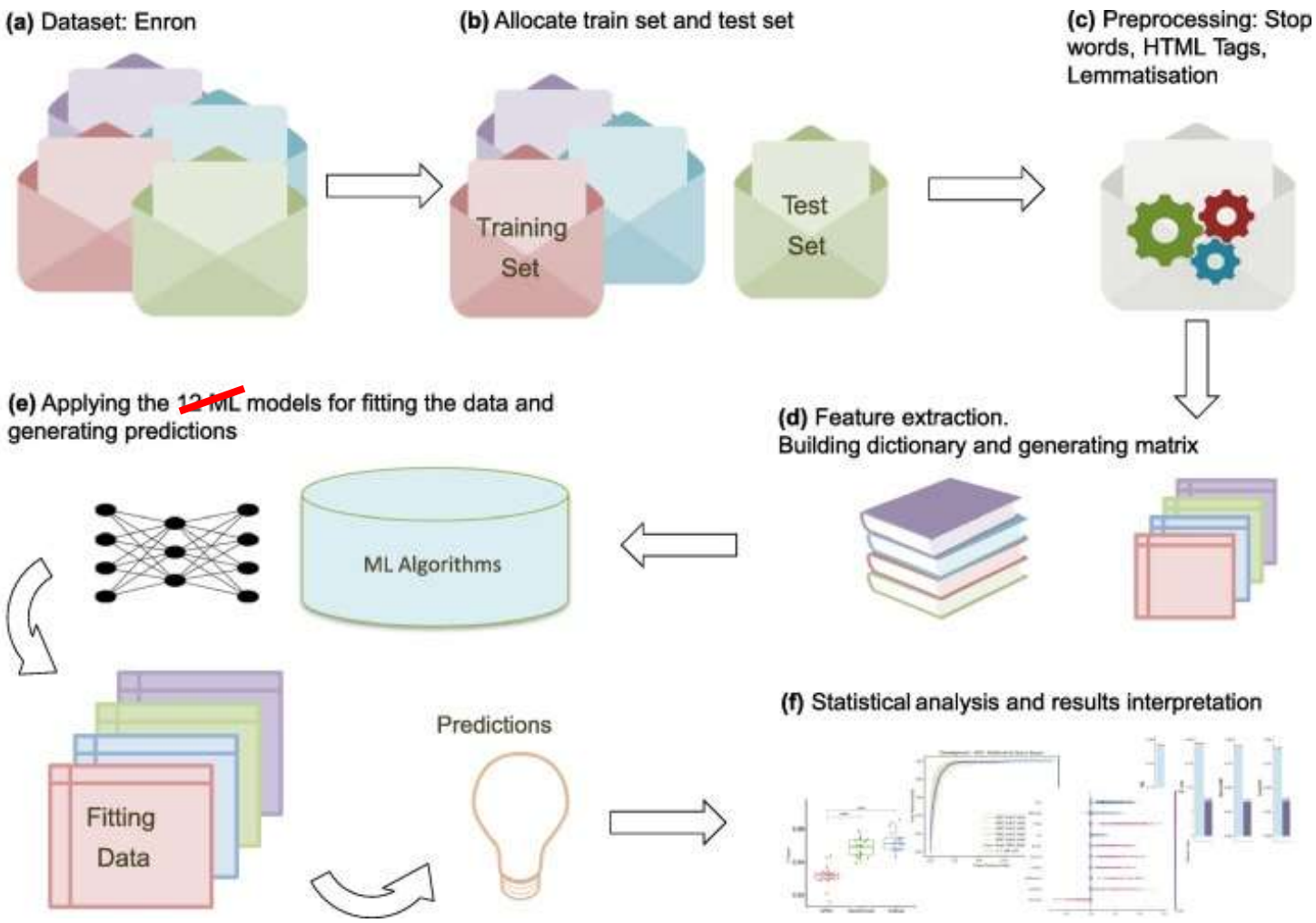


Fig. 1. **Pipeline diagram of the preprocessing and classification steps.** The pipeline shows the process applied for investigating the performance of 12 classifiers. (a) The Enron dataset is selected for the study; (b) the train and test sets are allocated (70% train set and 30% test set); (c) the preprocessing stage is then applied, including removal of stop words, HTML tag and lemmatisation; (d) the features are then extracted for generating a dictionary, based on the number of most occurring features; (e) finally, the 12 algorithms receive the matrices for fitting the data and predicting the classification outcomes; (f) comparative analysis is then performed to assess the statistical significance of the results and provide an accurate interpretation of the machine learning classification outcomes.

Model

Naïve Bayes

The most common machine learning algorithm for text classification.

Source: https://www.sciencedirect.com/science/article/pii/S0957417422005802

FOR THE PIPELINE

The recipes package that you can use to combine different feature engineering and preprocessing tasks into a single object and then apply these transformations to different data sets.

Source: https://www.tmwr.org/recipes

Join us!

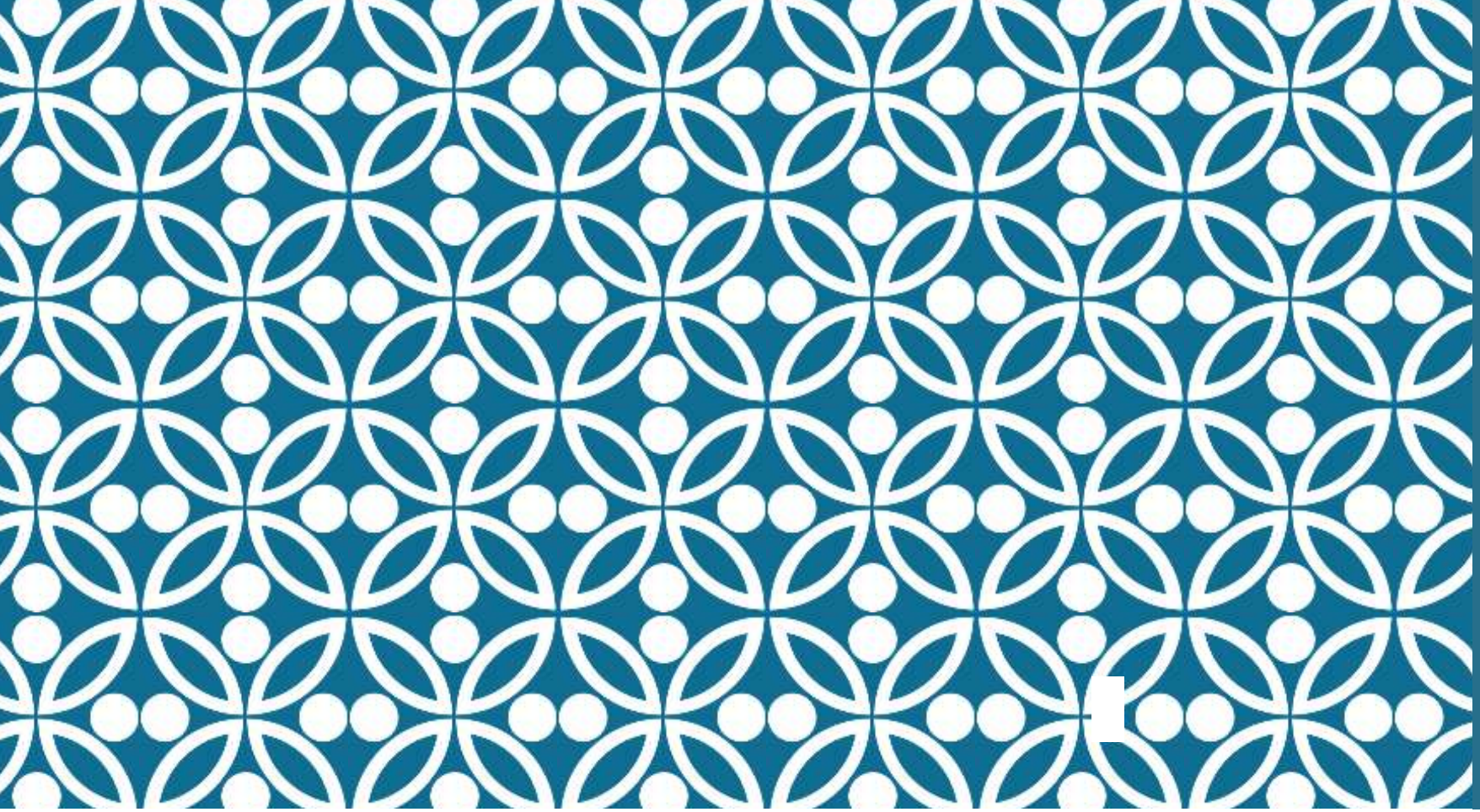
https://forms.office.com/e/8Bgd2YsasJ

All about the lab:

https://societal-analytics.nl/

Contact us at:

analytics-lab.fsw@vu.nl.



https://sofiag1l.github.io/

THANKS!

Dr. Sofia Gil-Clavel