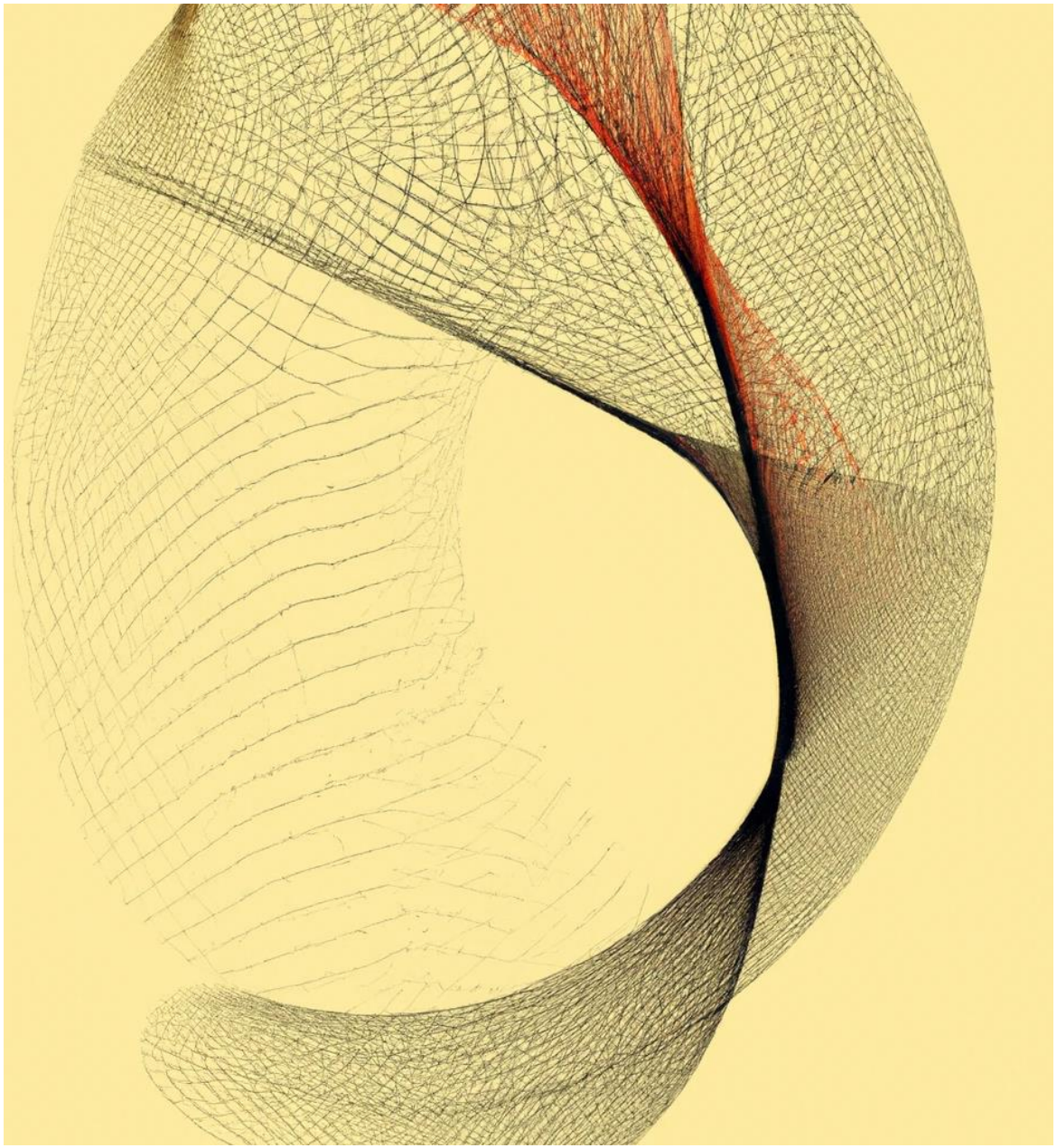
Introduction to Machine Learning in R

Dr Giorgia Mori

Dr Henry Lydecker

Sydney Informatic Hub

THE UNIVERSITY OF SYDNEY



Sydney Informatics Hub

“SIH is a Core Research Facility of the University of Sydney enabling excellence in computational and data-driven research through advanced digital infrastructure, expert data consultancy and analytics training”

Sydney Informatics Hub

Empowering researchers with modern data & computational methods

Research Computing

- High performance computing
- Bioinformatics & genomics
- Modelling & visualisation
- Computing training

Data Science & SWE

Consulting and project collaboration providing analysis & software development for data-driven research.

Statistical Consulting

- 1-on-1 consultancy for HDR-level and above.
- Experiment and survey design.
- Statistical model development and testing
- Statistics training

Consultation

Grant application support

Data Science, ML, modelling & analytics

Research compute platforms

Hacky Hour

Training

Code of Conduct

We expect all attendees of our training to follow the University of Sydney's Staff and Student Codes of Conduct, including the Bullying, harassment and discrimination prevention policy.

To foster a positive and professional learning environment, we encourage the following kinds of behaviours at all our events and platforms:

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Focus on what is best for the community
- Show courtesy and respect towards other community members

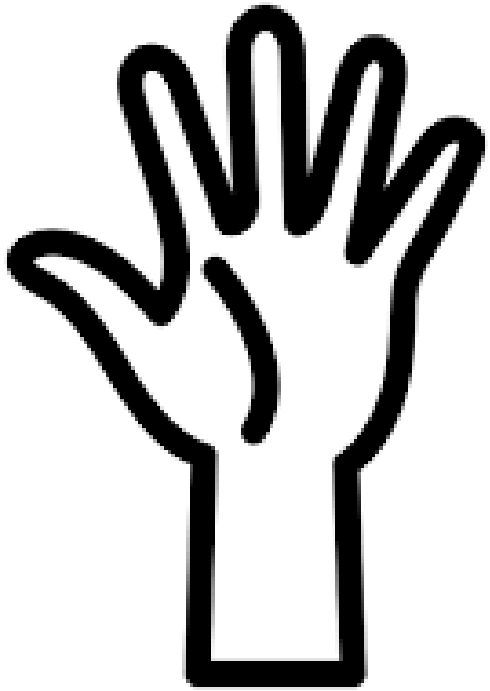
Our full CoC, with incident reporting guidelines, is available at
https://pages.github.sydney.edu.au/informatics/sih_codeofconduct/

(Examples of) Unacceptable behaviour

- Sustained disruption of talks, events or communications
- Written/verbal comments which have the effect of excluding people on the basis of membership of any specific group
- Causing someone to fear for their safety, such as through stalking, following, or intimidation
- Violent threats or language directed against another person
- The display of sexual or violent images
- Unwelcome sexual attention; non-consensual or unwelcome physical contact
- Insults or put downs; excessive swearing
- Sexist, racist, homophobic, transphobic, ableist, or exclusionary jokes
- Incitement to violence, suicide, or self-harm
- Continuing to initiate interaction (including photography or recording) with someone after being asked to stop
- Publication of private communication without consent

Asking for Help

I need help with my computer



I need help understanding something
(which likely means others do too)

Plan for the workshop

Time	Day1-March 28th	Day2-March 30th
9:00-9:15am	Welcome	Q&A from Day 1
9:15-10:30am	Intro to Machine Learning Exploratory Data Analysis (EDA) - Regression	Intro to Classification and EDA
10:45-11:00am	*Morning break*	*Morning break*
11:00am-12:30pm	EDA - Regression (continue)	Intro to Classification and EDA (continue)
12:30-1:30pm	*Lunch*	*Lunch*
1:30-3:00pm	Intro to Tidymodels – pre-processing	Tuning hyperparameters
3:00-3:15pm	*Afternoon break*	*Afternoon break*
3:15-4:45pm	Linear Models and working with workflows	Regularized logistic regression, random forest – compare models

How are statistics and Machine Learning related?

Statistics: inferring from samples

Machine Learning: making generalizable predictors

- 5 beds
- 4 baths
- Water views
- Mosman
- Featured on fancy architecture website

Skill comes from learning and practice

Terminology

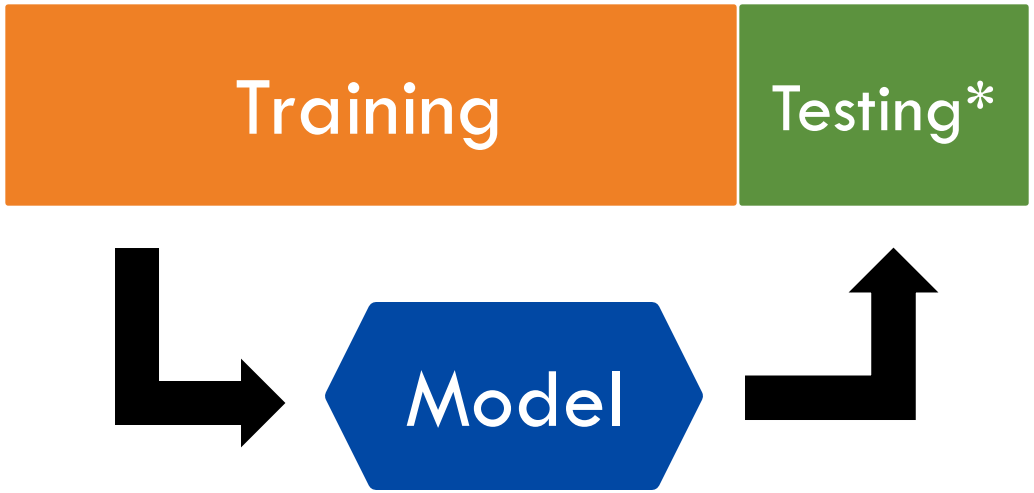
- Find an **algorithm** f(x) that most accurately predicts future values y based on a set of inputs x
- Observations/data points – rows (in R)
- Predictor variables – columns (in R)
- Outcome - what you're trying to predict (usually more applicable in the context of supervised learning)

Generalisability and overfitting

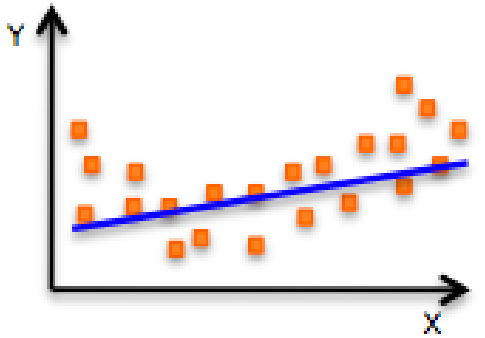
- **Generalisability** – ability to predict well on future, unseen data (can assess this with a validation set) – **good model!**

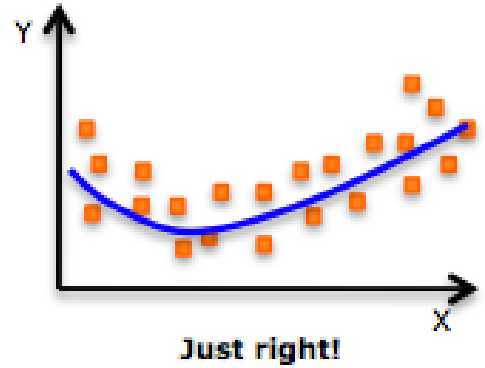
- **Overfitting** – model fits existing data very well *(too well)*, but doesn't generalise to new data – **bad model!**

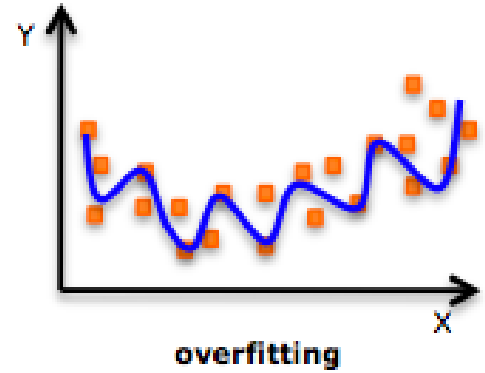
Training and testing



We want to train our model such that it can be generalizable enough to the test set, but not such that it is overfit to the training set.

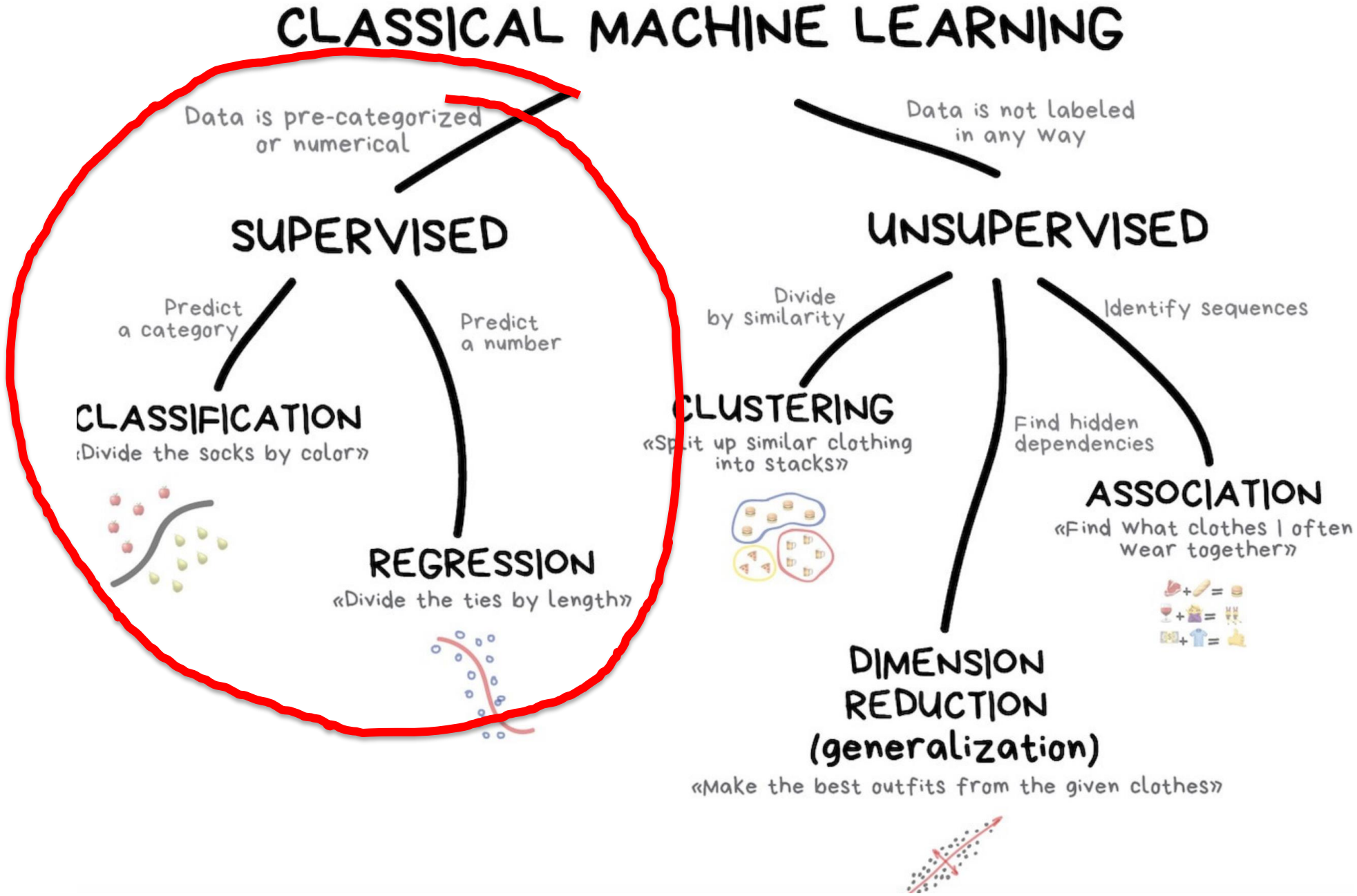






Images from: *http://epicalsoft.blogspot.com/2019/02/azure-machine-learning-sobreajuste-y.html*

***Do not touch the test set!**



https://vas3k.com/blog/machine_learning/

Your turn!

- Form teams within your table;
- Share your backgrounds with R, data, and Machine Learning;
- Choose a team name.

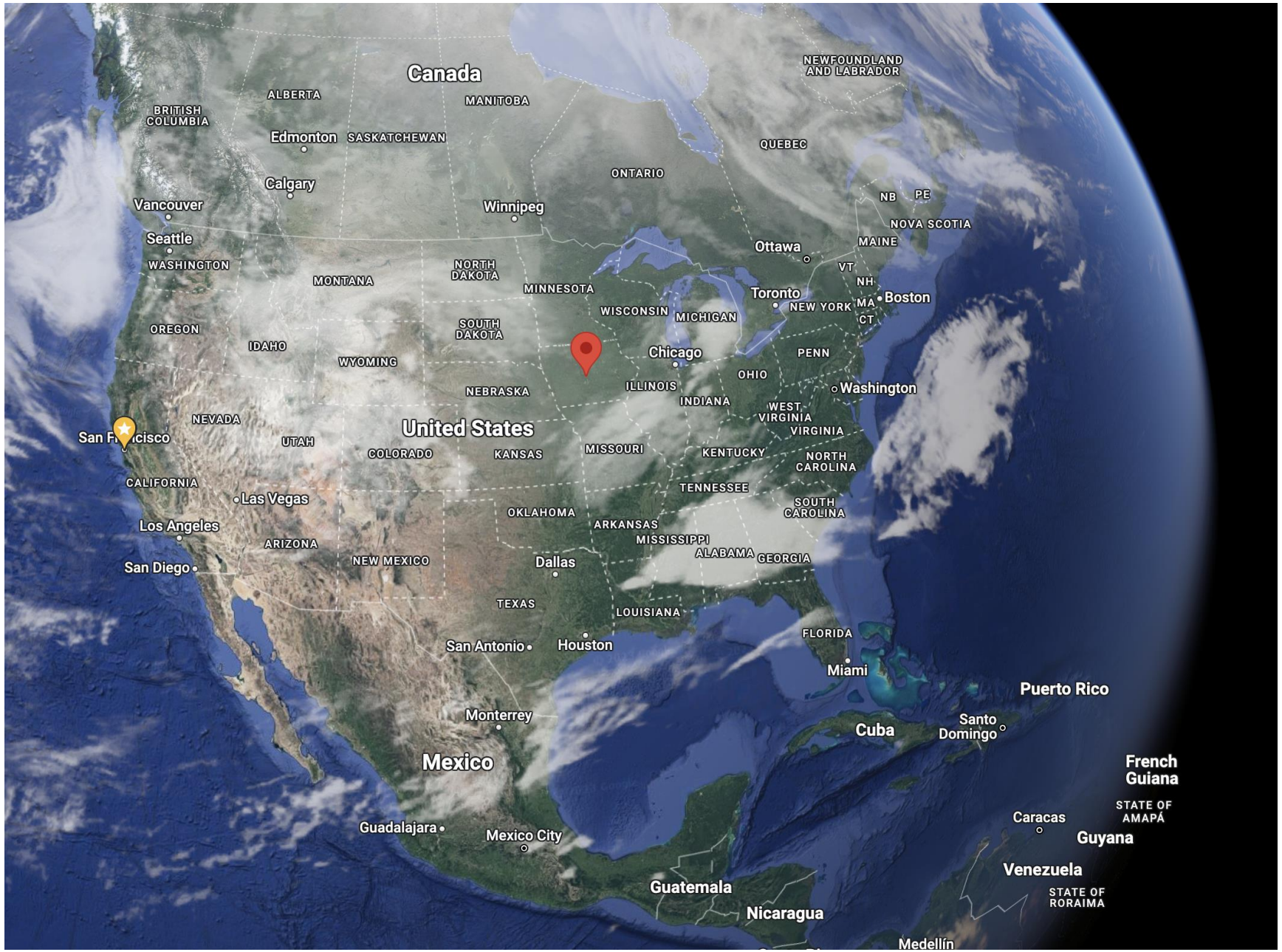
Dataset - AmesHousing

Descriptions of 2,930 houses sold in Ames, IA from 2006 to 2010, collected by the Ames Assessor's Office.

```
#install.packages("AmesHousing")
library(AmesHousing)
ameshousing <- AmesHousing::make_ames()
```

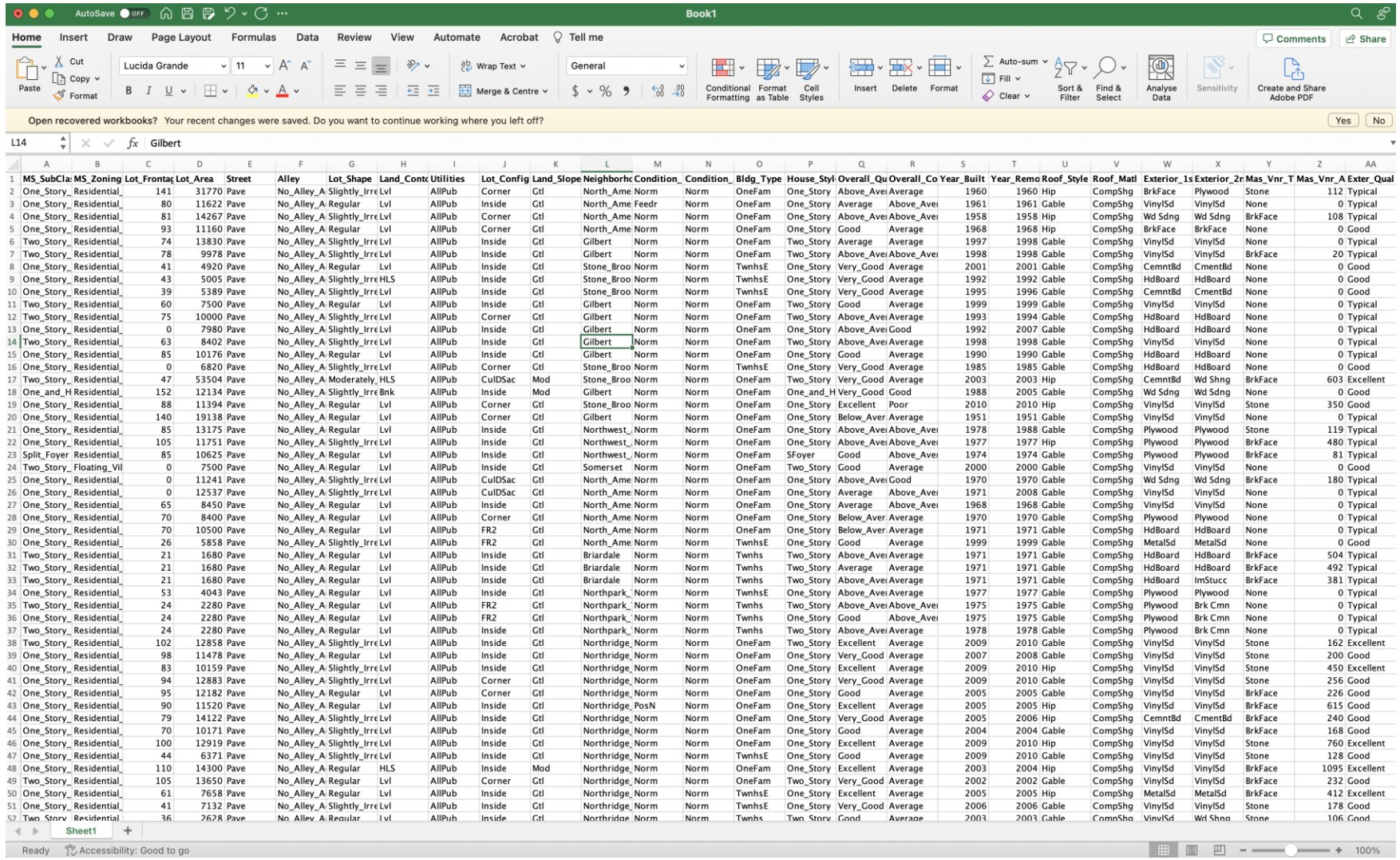
Meet the data: Ames housing

- Data set of housing data from Ames, Iowa, collected from 2006 – 2010.
- Population of ~58,000 (2010)
- Data includes many different physical measurements or descriptions of properties.
- Assembled for an end of semester regression project by Dean De Cock at Truman University.
- Popularized by a Kaggle competition predicting housing prices


Canada
NEWFOUNDLAND AND LABRADOR
ALBERTA
MANITOBA
BRITISH COLUMBIA
Edmonton
SASKATCHEWAN
QUEBEC
ONTARIO
Calgary
Vancouver
Winnipeg
NB
PE
NOVA SCOTIA
Seattle
Ottawa
MAINE
WASHINGTON
NORTH DAKOTA
VT
MONTANA
MINNESOTA
NH
Toronto
Boston
NEW YORK
MA
WISCONSIN
MICHIGAN
CT
OREGON
SOUTH DAKOTA
IDAHO
Chicago
PENN
WYOMING
OHIO
NEBRASKA
ILLINOIS
Washington
INDIANA
WEST VIRGINIA
NEVADA
United States
VIRGINIA
San Francisco
UTAH
COLORADO
KANSAS
MISSOURI
KENTUCKY
NORTH CAROLINA
CALIFORNIA
TENNESSEE
Las Vegas
SOUTH CAROLINA
OKLAHOMA
Los Angeles
ARKANSAS
MISSISSIPPI
ARIZONA
ALABAMA
GEORGIA
San Diego
NEW MEXICO
Dallas
TEXAS
LOUISIANA
FLORIDA
San Antonio
Houston
Miami
Puerto Rico
Monterrey
Cuba
Santo Domingo
Mexico
French Guiana
STATE OF AMAPÁ
Caracas
Guadalajara
Mexico City
Guyana
Venezuela
Guatemala
STATE OF RORAIMA
Nicaragua
Medellín

Exploratory Data Analysis (EDA)

Exploring Our Data



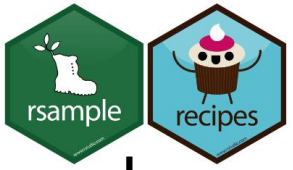
Tidymodels

Your data budget - Data splitting

For machine learning, we typically split data into training and test sets:

- The **training set** is used to estimate model parameters.
- The **test set** is used to find an independent assessment of model performance.

Your data budget - Data spending

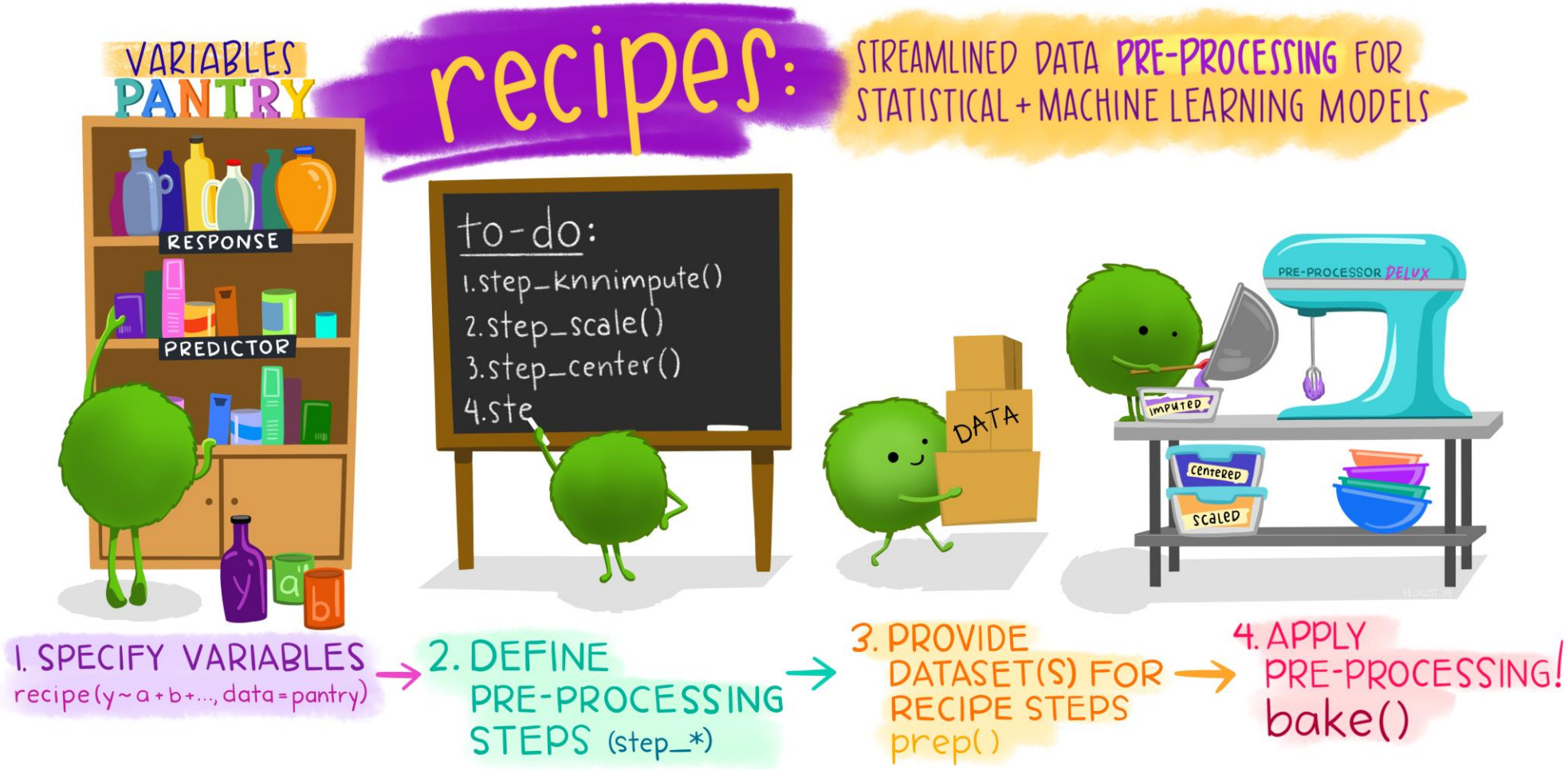


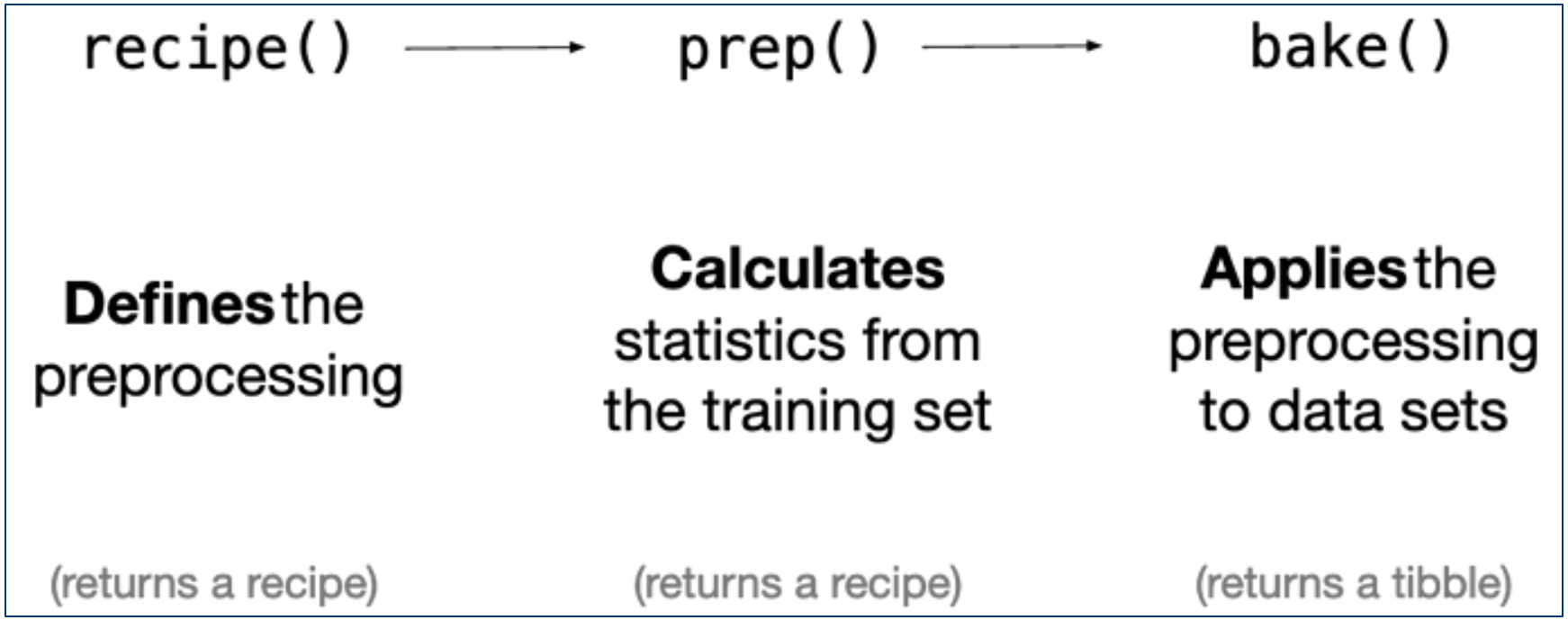
- Spending too much data in training prevents us from computing a good assessment of predictive performance;

- Spending too much data in testing prevents us from computing a good estimate of model parameters.

Some commonly used cut-offs include:
•60% training / 40% testing
•70% training / 30% testing
•80% training / 20% testing

Prepare your data for modeling

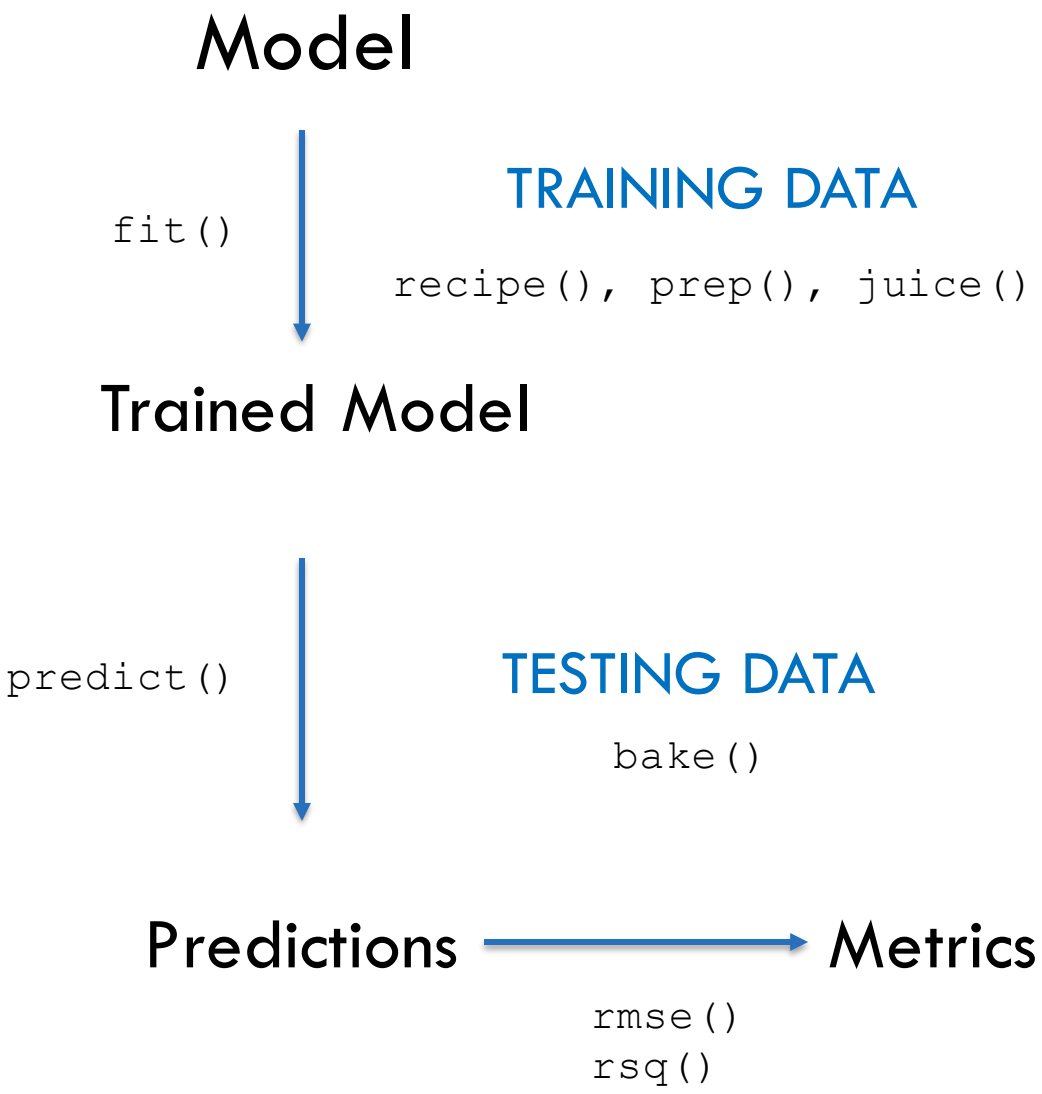




Within `recipe()`, the **training set** is used to determine the data types of each column

Within `prep()`, the preprocessing steps are executed using the **training set**

Within `bake()`, the **training parameters** are applied to the **testing set.**


Model
fit()
TRAINING DATA
recipe(), prep(), juice()
Trained Model
predict()
TESTING DATA
bake()
Predictions
Metrics
rmse()
rsq()

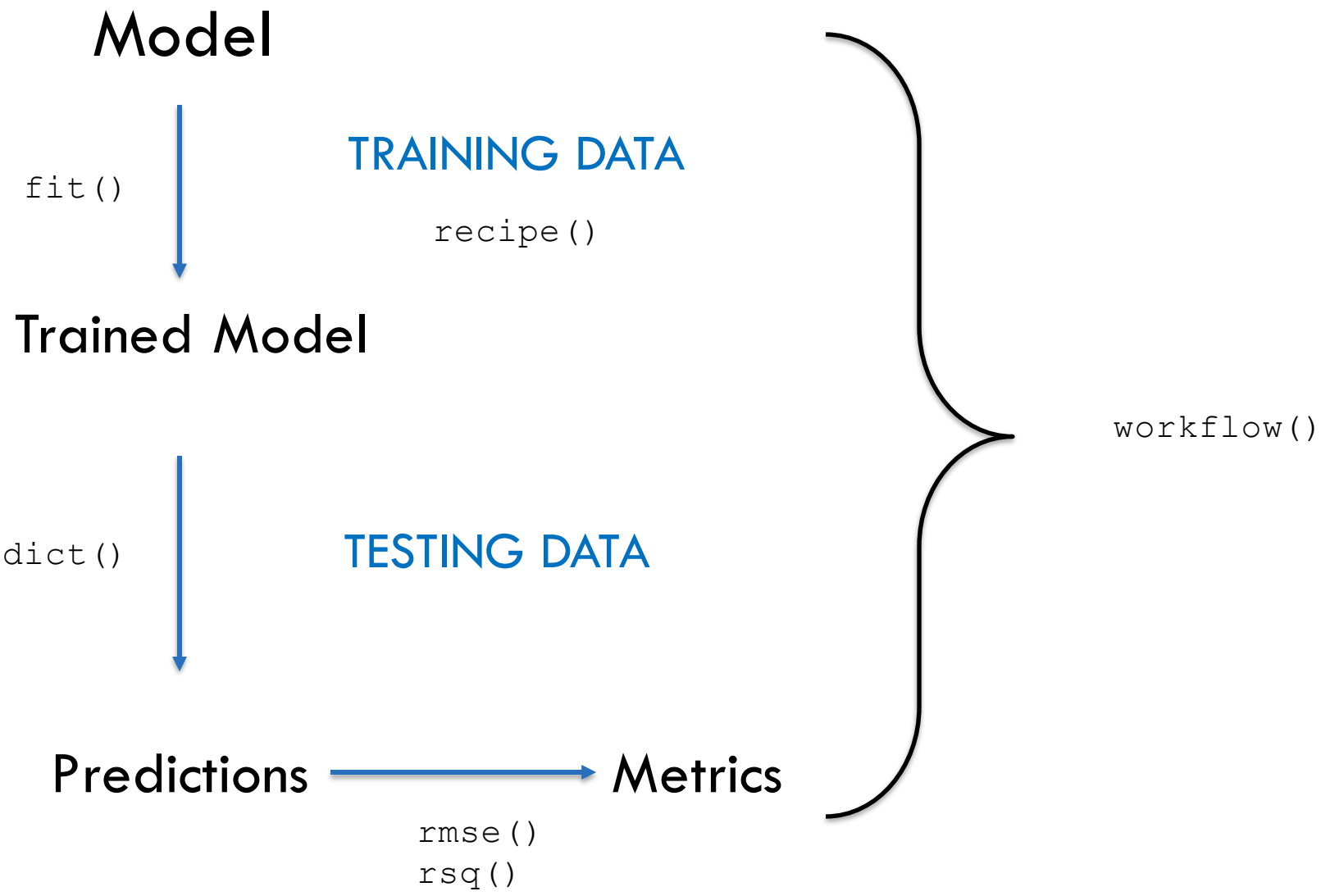
Let’s make our life easier!

Workflows

- The recipe **prepping** and **model fitting** can be executed using a single call to *fit()* instead of *prep()-juice()-fit()*;

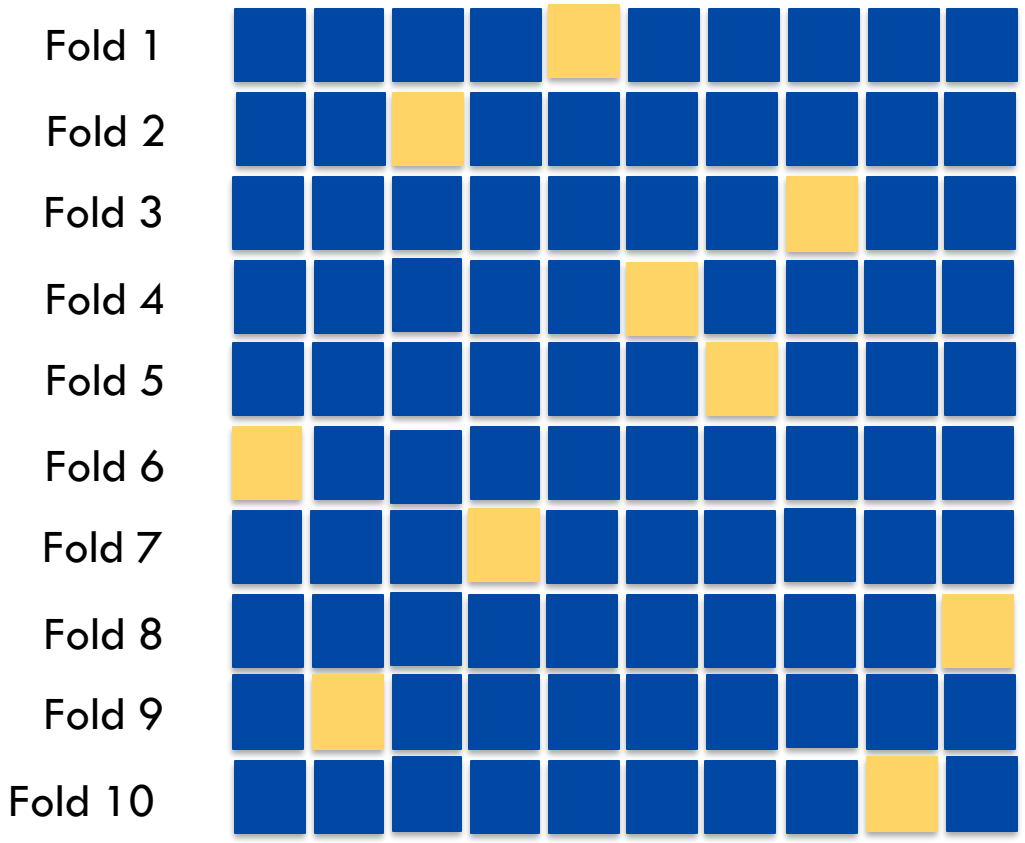
- The recipe **baking** and **model predictions** are handled with a single call to *predict()* instead of *bake()-predict()*.

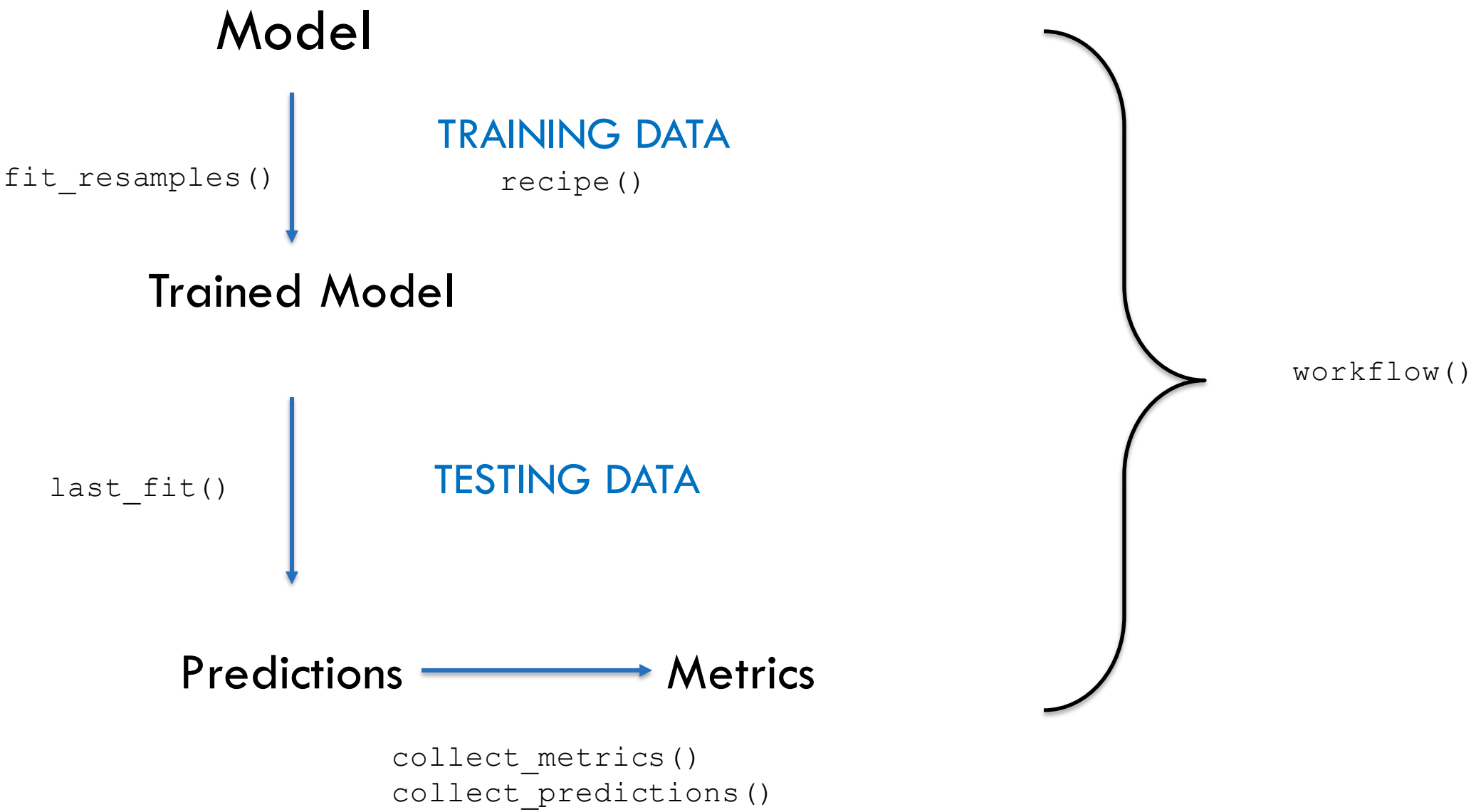
Within `fit()`, the **training** data are used for all estimation operations (from the recipe that is part of the `workflow()`)



Within `predict()`, values from the **training** data are used to standardize the **testing** data.

Resampling – cross-validation




Model
fit_resamples()
TRAINING DATA
recipe()
Trained Model
last_fit()
TESTING DATA
Predictions
Metrics
collect_metrics()
collect_predictions()
workflow()

Resources to keep learning:

- https://www.tidymodels.org/
- https://www.tmwr.org/
- http://www.feat.engineering/
- https://smltar.com/

Your feedback is important!