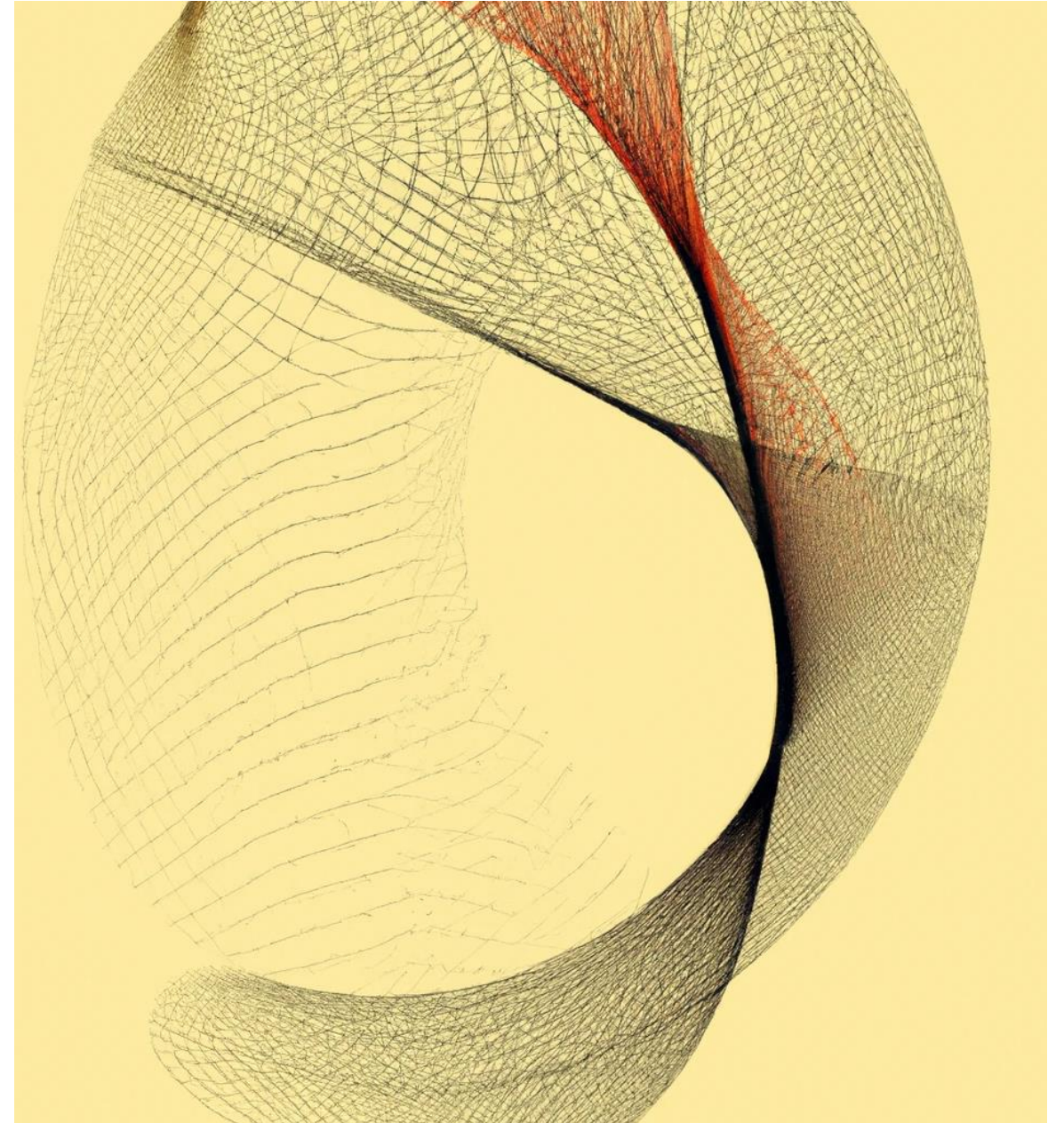


Introduction to Machine Learning in R

Dr Giorgia Mori

Dr Henry Lydecker

Sydney Informatic Hub



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:

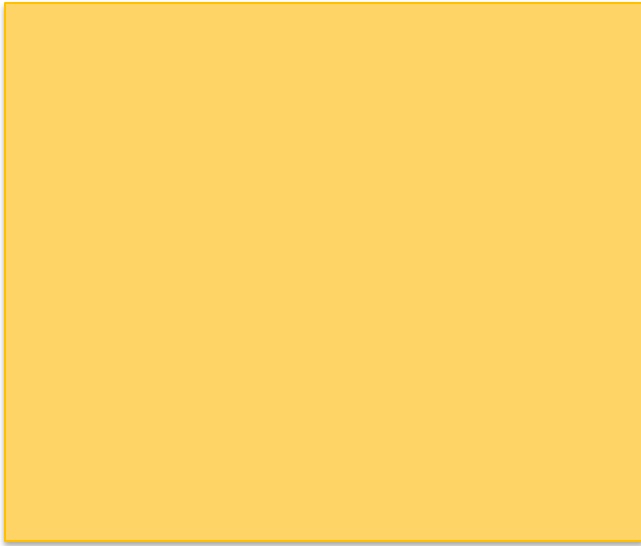
- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Focus on what is best for the community
- Show courtesy and respect towards other community members

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 28 th	Day2-March 30 th
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	<i>Morning break</i>	<i>Morning break</i>
11:00am-12:30pm	EDA - Regression (continue)	Intro to Classification and EDA (continue)
12:30-1:30pm	<i>Lunch</i>	<i>Lunch</i>
1:30-3:00pm	Intro to Tidymodels – pre-processing	Tuning hyperparameters
3:00-3:15pm	<i>Afternoon break</i>	<i>Afternoon break</i>
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



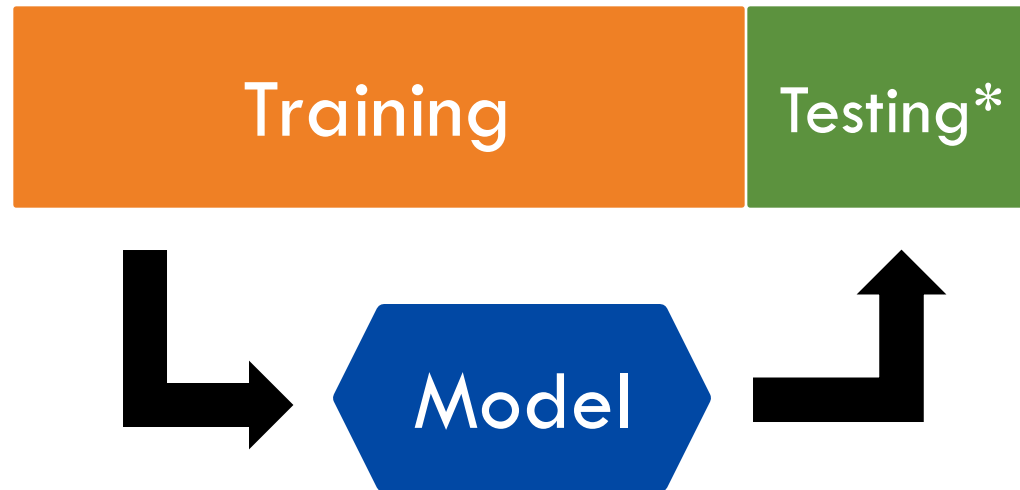
Terminology

- 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)

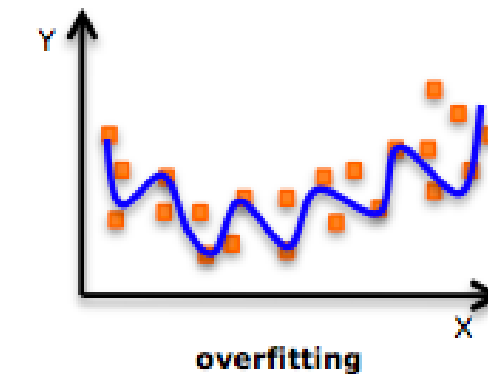
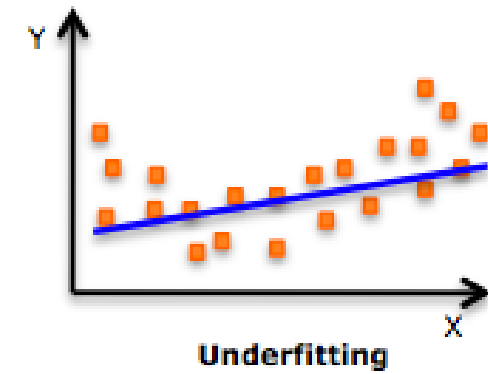
Generalisability and overfitting

- **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

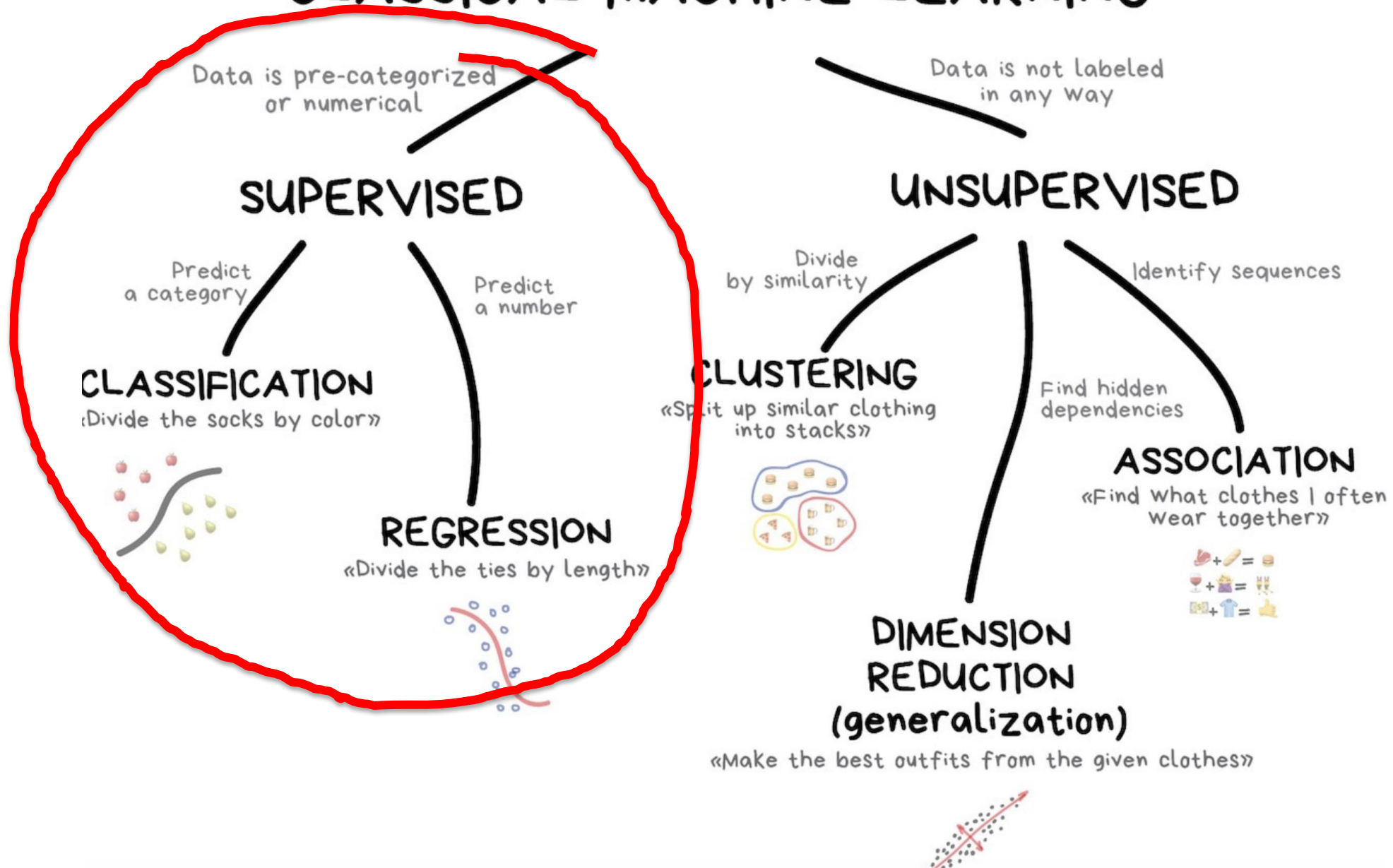


We want to train our model such that it can be generalizable enough to the test set, but not such that it is overfit to the training set.



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

CLASSICAL MACHINE LEARNING



Your turn!

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

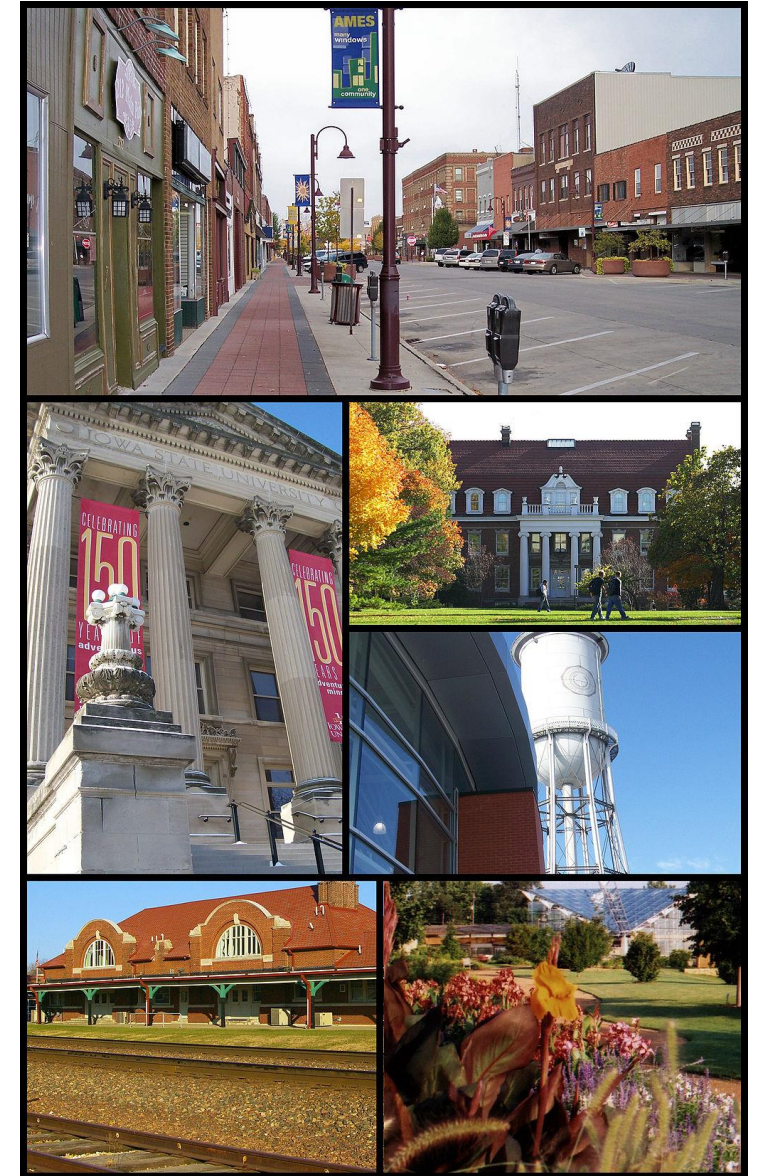
Dataset - AmesHousing

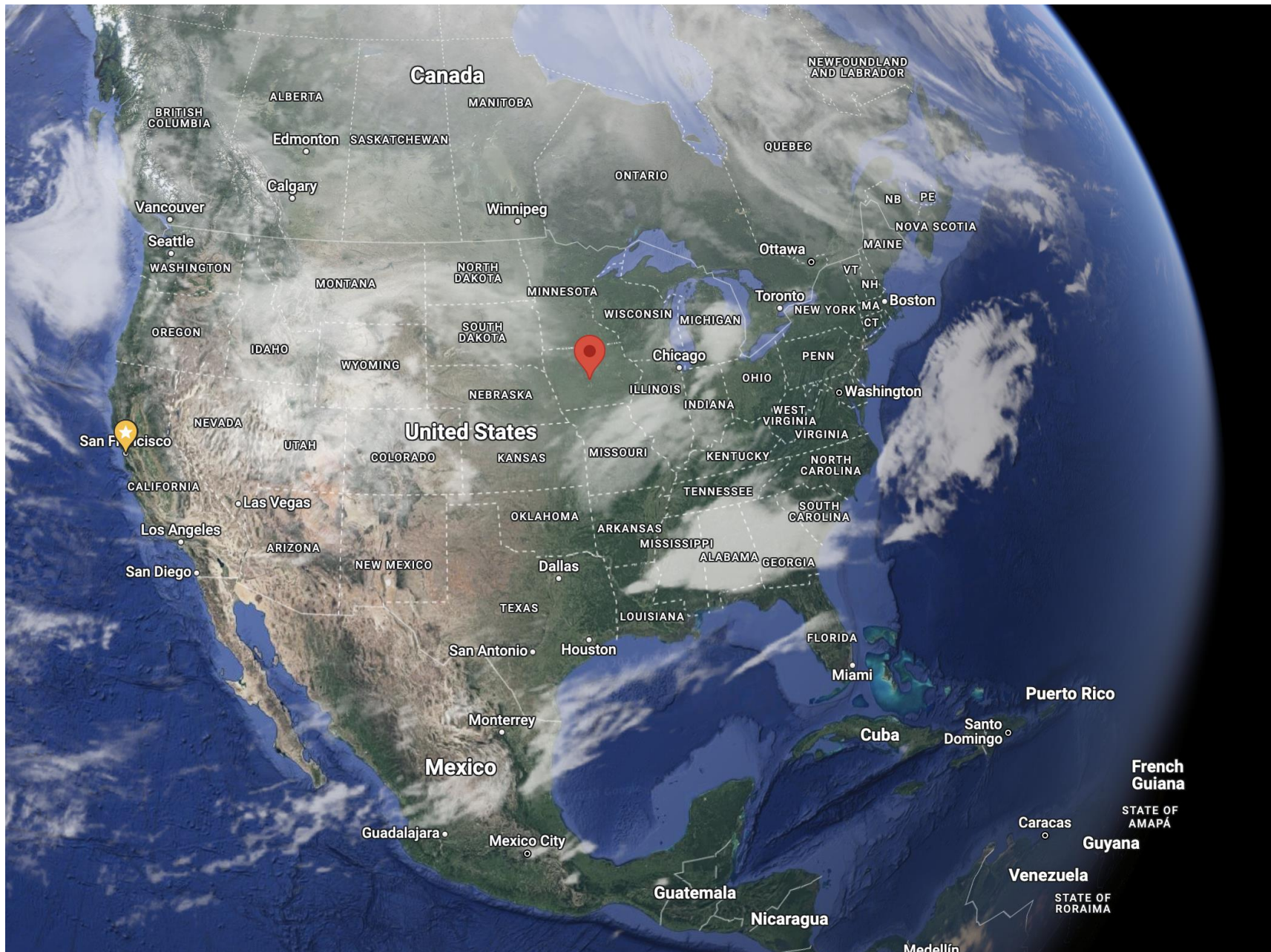
Descriptions of 2,930 houses sold in Ames, IA from 2006 to 2010, collected by the Ames Assessor's Office.

```
#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





Exploratory Data Analysis (EDA)

Exploring Our Data

AutoSave OFF

Book1

HomeInsertDrawPage LayoutFormulasDataReviewViewAutomateAcrobatTell me

PasteCutCopyFormat

Lucida Grande11A

B**I**U

Wrap Text

General

Conditional FormattingFormat as TableCell Styles

InsertDeleteFormat

Auto-sumFillClear

Sort & FilterFind & Select

Analyse DataSensitivity

Create and ShareAdobe PDF

CommentsShare

Open recovered workbooks? Your recent changes were saved. Do you want to continue working where you left off?

YesNo

L14X✓fxGilbert

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA		
1	MS_SubCia:	MS_Zoning	Lot_Fronta	Lot_Area	Street	Alley	Lot_Shape	Land_Conc	Utilities	Lot_Config	Land_Slope	Neighborch	Condition	Condition	Bldg_Type	House_Stat	Overall_Qu	Overall_Co	Year_Built	Year_Remo	Roof_Style	Roof_Mat	Exterior_1s	Exterior_2r	Mas_Vnr_T	Mas_Vnr_A	Exter_Qual		
2	One_Story_Residential		141	31770	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Corner	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Above_Ave	Average	1960	1960	Hip	CompShg	BrkFace	Plywood	Stone		112	Typical		
3	One_Story_Residential		80	11622	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	North_Ame	Feedr	Norm	OneFam	One_Story	Average	Above_Ave	1961	1961	Gable	CompShg	VinylSd	VinylSd	None		0	Typical	
4	One_Story_Residential		81	14267	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Corner	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Above_Ave	Above_Ave	1958	1958	Hip	CompShg	Wd Sdng	Wd Sdng	BrkFace		108	Typical		
5	One_Story_Residential		93	11160	Pave	No_Alley_A	Regular	Lvl	AllPub	Corner	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Good	Average	1968	1968	Hip	CompShg	BrkFace	BrkFace	None		0	Good	
6	Two_Story_Residential		74	13830	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	OneFam	Two_Story	Average	Average	1997	1998	Gable	CompShg	VinylSd	VinylSd	None		0	Typical		
7	Two_Story_Residential		78	9978	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	OneFam	Two_Story	Above_Ave	Above_Ave	1998	1998	Gable	CompShg	VinylSd	VinylSd	BrkFace		20	Typical		
8	One_Story_Residential		41	4920	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Stone_Broo	Norm	Norm	TwmhsE	One_Story	Very_Good	Average	2001	2001	Gable	CompShg	CemntBd	CemntBd	None		0	Good	
9	One_Story_Residential		43	5005	Pave	No_Alley_A	Slightly_IrreHLS	AllPub	Inside	Gtl	Stone_Broo	Norm	Norm	Norm	TwmhsE	One_Story	Very_Good	Average	1992	1992	Gable	CompShg	HdBoard	HdBoard	None		0	Good	
10	One_Story_Residential		39	5389	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Stone_Broo	Norm	Norm	Norm	TwmhsE	One_Story	Very_Good	Average	1995	1996	Gable	CompShg	CemntBd	CemntBd	None		0	Good	
11	Two_Story_Residential		60	7500	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	OneFam	Two_Story	Good	Average	1999	1999	Gable	CompShg	VinylSd	VinylSd	None		0	Typical	
12	Two_Story_Residential		75	10000	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	OneFam	Two_Story	Above_Ave	Average	1993	1994	Gable	CompShg	HdBoard	HdBoard	None		0	Typical		
13	One_Story_Residential		0	7980	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	OneFam	One_Story	Above_Ave	Good	1992	2007	Gable	CompShg	HdBoard	HdBoard	None		0	Typical		
14	Two_Story_Residential		63	8402	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	OneFam	Two_Story	Above_Ave	Average	1998	1998	Gable	CompShg	VinylSd	VinylSd	None		0	Typical		
15	One_Story_Residential		85	10176	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	OneFam	One_Story	Good	Average	1990	1990	Gable	CompShg	HdBoard	HdBoard	None		0	Typical	
16	One_Story_Residential		0	6820	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Corner	Gtl	Stone_Broo	Norm	Norm	Norm	TwmhsE	One_Story	Very_Good	Average	1985	1985	Gable	CompShg	HdBoard	HdBoard	None		0	Good	
17	Two_Story_Residential		47	53504	Pave	No_Alley_A	Moderately_HLS	AllPub	CulDSac	Mod	Stone_Broo	Norm	Norm	OneFam	Two_Story	Very_Good	Average	2003	2003	Hip	CompShg	CemntBd	Wd Sdng	BrkFace		603	Excellent		
18	One_and_H	Residential	152	12134	Pave	No_Alley_A	Slightly_IrreBnk	AllPub	Inside	Mod	Gilbert	Norm	Norm	OneFam	One_and_H	Very_Good	Good	1988	2005	Gable	CompShg	Wd Sdng	Wd Sdng	None		0	Good		
19	One_Story_Residential		88	11394	Pave	No_Alley_A	Regular	Lvl	AllPub	Corner	Gtl	Stone_Broo	Norm	Norm	OneFam	One_Story	Excellent	Poor	2010	2010	Hip	CompShg	VinylSd	VinylSd	Stone		350	Good	
20	One_Story_Residential		140	19138	Pave	No_Alley_A	Regular	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	OneFam	One_Story	Below_Ave	Average	1951	1951	Gable	CompShg	VinylSd	VinylSd	None		0	Typical	
21	One_Story_Residential		85	13175	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northwest	Norm	Norm	OneFam	One_Story	Above_Ave	Above_Ave	1978	1988	Gable	CompShg	Plywood	Plywood	Stone		119	Typical	
22	One_Story_Residential		105	11751	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northwest	Norm	Norm	OneFam	One_Story	Above_Ave	Above_Ave	1977	1977	Hip	CompShg	Plywood	Plywood	BrkFace		480	Typical		
23	Split_Foyer	Residential	85	10625	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northwest	Norm	Norm	OneFam	SFoyer	Good	Above_Ave	1974	1974	Gable	CompShg	Plywood	Plywood	BrkFace		81	Typical	
24	Two_Story_Floating_Vil		0	7500	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Somerset	Norm	Norm	OneFam	Two_Story	Good	Average	2000	2000	Gable	CompShg	VinylSd	VinylSd	None		0	Good	
25	One_Story_Residential		0	11241	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	CulDSac	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Above_Ave	Good	1970	1970	Gable	CompShg	Wd Sdng	Wd Sdng	BrkFace		180	Typical		
26	One_Story_Residential		0	12537	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	CulDSac	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Average	Above_Ave	1971	2008	Gable	CompShg	VinylSd	VinylSd	None		0	Typical		
27	One_Story_Residential		65	8450	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Average	Above_Ave	1968	1968	Gable	CompShg	VinylSd	VinylSd	None		0	Typical	
28	One_Story_Residential		70	8400	Pave	No_Alley_A	Regular	Lvl	AllPub	Corner	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Below_Ave	Average	1970	1970	Gable	CompShg	Plywood	Plywood	None		0	Typical	
29	One_Story_Residential		70	10500	Pave	No_Alley_A	Regular	Lvl	AllPub	FR2	Gtl	North_Ame	Norm	Norm	OneFam	One_Story	Below_Ave	Average	1971	1971	Gable	CompShg	HdBoard	HdBoard	None		0	Typical	
30	One_Story_Residential		26	5858	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	FR2	Gtl	North_Ame	Norm	Norm	Norm	TwmhsE	One_Story	Good	Average	1999	1999	Gable	CompShg	MetalSd	MetalSd	None		0	Good	
31	Two_Story_Residential		21	1680	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Brriardale	Norm	Norm	Twmhs	Two_Story	Above_Ave	Average	1971	1971	Gable	CompShg	HdBoard	HdBoard	BrkFace		504	Typical	
32	Two_Story_Residential		21	1680	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Brriardale	Norm	Norm	Twmhs	Two_Story	Average	Average	1971	1971	Gable	CompShg	HdBoard	HdBoard	BrkFace		492	Typical	
33	Two_Story_Residential		21	1680	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Brriardale	Norm	Norm	Twmhs	Two_Story	Above_Ave	Average	1971	1971	Gable	CompShg	HdBoard	ImStucc	BrkFace		381	Typical	
34	One_Story_Residential		53	4043	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northpark	Norm	Norm	TwmhsE	One_Story	Above_Ave	Average	1977	1977	Gable	CompShg	Plywood	Plywood	None		0	Typical	
35	Two_Story_Residential		24	2280	Pave	No_Alley_A	Regular	Lvl	AllPub	FR2	Gtl	Northpark	Norm	Norm	Twmhs	Two_Story	Above_Ave	Above_Ave	1975	1975	Gable	CompShg	Plywood	Brk Cmn	None		0	Typical	
36	One_Story_Residential		24	2280	Pave	No_Alley_A	Regular	Lvl	AllPub	FR2	Gtl	Northpark	Norm	Norm	Twmhs	One_Story	Good	Above_Ave	1975	1975	Gable	CompShg	Plywood	Brk Cmn	None		0	Typical	
37	Two_Story_Residential		24	2280	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northpark	Norm	Norm	Twmhs	Two_Story	Above_Ave	Average	1978	1978	Gable	CompShg	Plywood	Brk Cmn	None		0	Typical	
38	Two_Story_Residential		102	12858	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	OneFam	Two_Story	Excellent	Average	2009	2010	Gable	CompShg	VinylSd	VinylSd	Stone		162	Excellent		
39	One_Story_Residential		98	11478	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Very_Good	Average	2007	2008	Gable	CompShg	VinylSd	VinylSd	Stone		200	Good	
40	One_Story_Residential		83	10159	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Excellent	Average	2009	2010	Hip	CompShg	VinylSd	VinylSd	Stone		450	Excellent		
41	One_Story_Residential		94	12883	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Corner	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Very_Good	Average	2009	2010	Gable	CompShg	VinylSd	VinylSd	Stone		256	Good		
42	One_Story_Residential		95	12182	Pave	No_Alley_A	Regular	Lvl	AllPub	Corner	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Good	Average	2005	2005	Gable	CompShg	VinylSd	VinylSd	BrkFace		226	Good	
43	One_Story_Residential		90	11520	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northridge_PosN	Norm	Norm	OneFam	One_Story	Excellent	Average	2005	2005	Hip	CompShg	VinylSd	VinylSd	BrkFace		615	Good	
44	One_Story_Residential		79	14122	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Very_Good	Average	2005	2006	Hip	CompShg	CemntBd	CemntBd	BrkFace		240	Good		
45	One_Story_Residential		70	10171	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Good	Average	2004	2004	Gable	CompShg	VinylSd	VinylSd	BrkFace		168	Good		
46	One_Story_Residential		100	12919	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	OneFam	One_Story	Excellent	Average	2009	2010	Hip	CompShg	VinylSd	VinylSd	Stone		760	Excellent		
47	One_Story_Residential		44	6371	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	Norm	TwmhsE	One_Story	Good	Average	2009	2010	Gable	CompShg	VinylSd	VinylSd	Stone		128	Good	
48	One_Story_Residential		110	14300	Pave	No_Alley_A	Regular	HLS	AllPub	Inside	Mod	Northridge	Norm	Norm	OneFam	One_Story	Excellent	Average	2003	2004	Hip	CompShg	VinylSd	VinylSd	BrkFace		1095	Excellent	
49	Two_Story_Residential		105	13650	Pave	No_Alley_A	Regular	Lvl	AllPub	Corner	Gtl	Northridge	Norm	Norm	OneFam	Two_Story	Very_Good	Average	2002	2002	Gable	CompShg	VinylSd	VinylSd	BrkFace		232	Good	
50	One_Story_Residential		61	7658	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	Norm	TwmhsE	One_Story	Excellent	Average	2005	2005	Hip	CompShg	MetalSd	MetalSd	BrkFace		412	Excellent
51	One_Story_Residential		41	7132	Pave	No_Alley_A	Slightly_IrreLvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	Norm	TwmhsE	One_Story	Very_Good	Average	2006	2006	Gable	CompShg	VinylSd	VinylSd	Stone		178	Good	
52	Two_Story_Residential		36	2628	Pave	No_Alley_A	Regular	Lvl	AllPub	Inside	Gtl	Northridge	Norm	Norm	Norm	Twmhs	Two_Story	Good	Average	2003	2003	Gable	CompShg	VinylSd	Wd Sdng	Stone		106	Good

Sheet1+

ReadyAccessibility: Good to go

100%

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.

Original
Training
Testing



Your data budget - Data spending



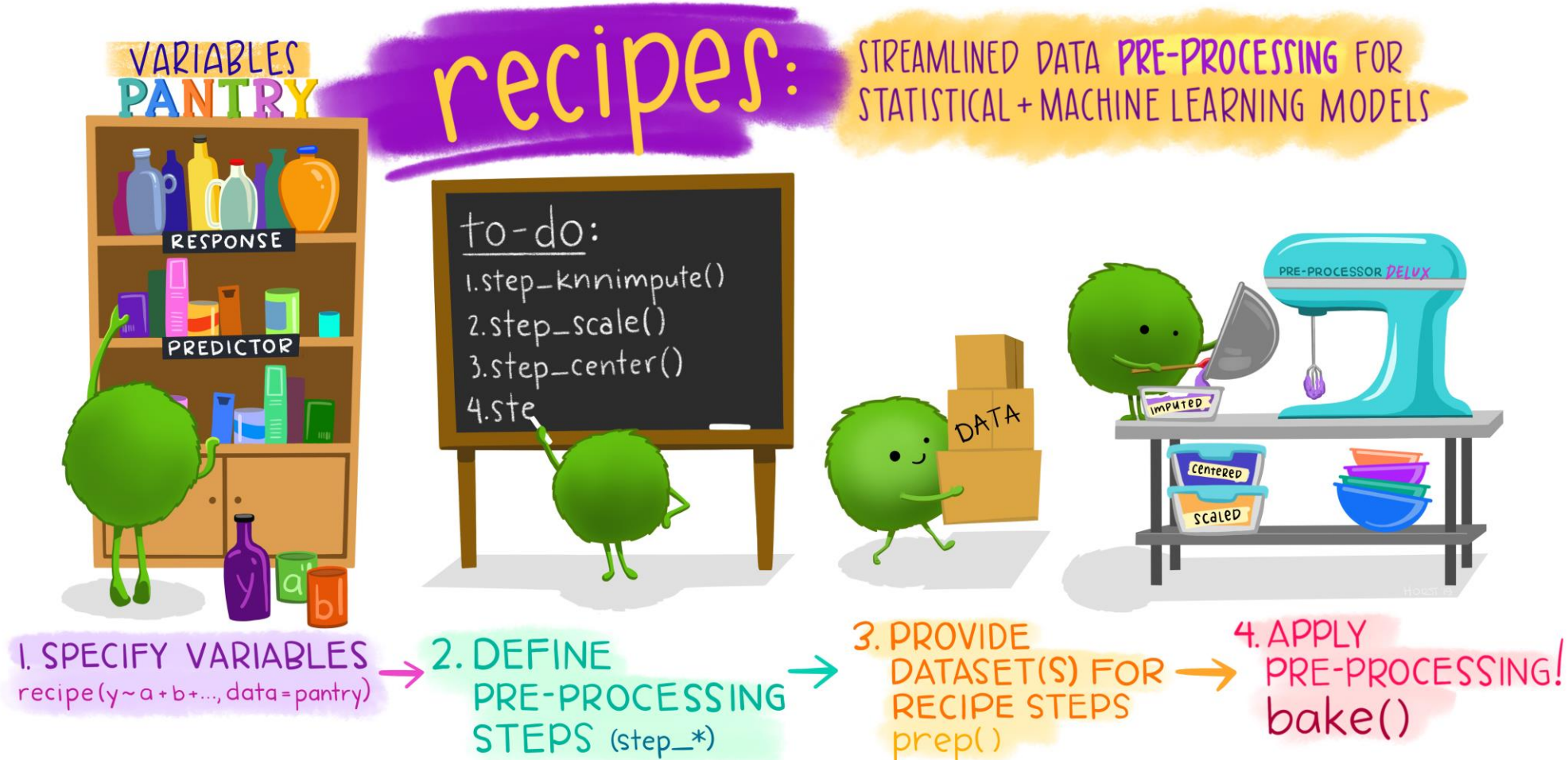
- Spending too much data in training prevents us from computing a good assessment of predictive performance;
- Spending too much data in testing prevents us from computing a good estimate of model parameters.

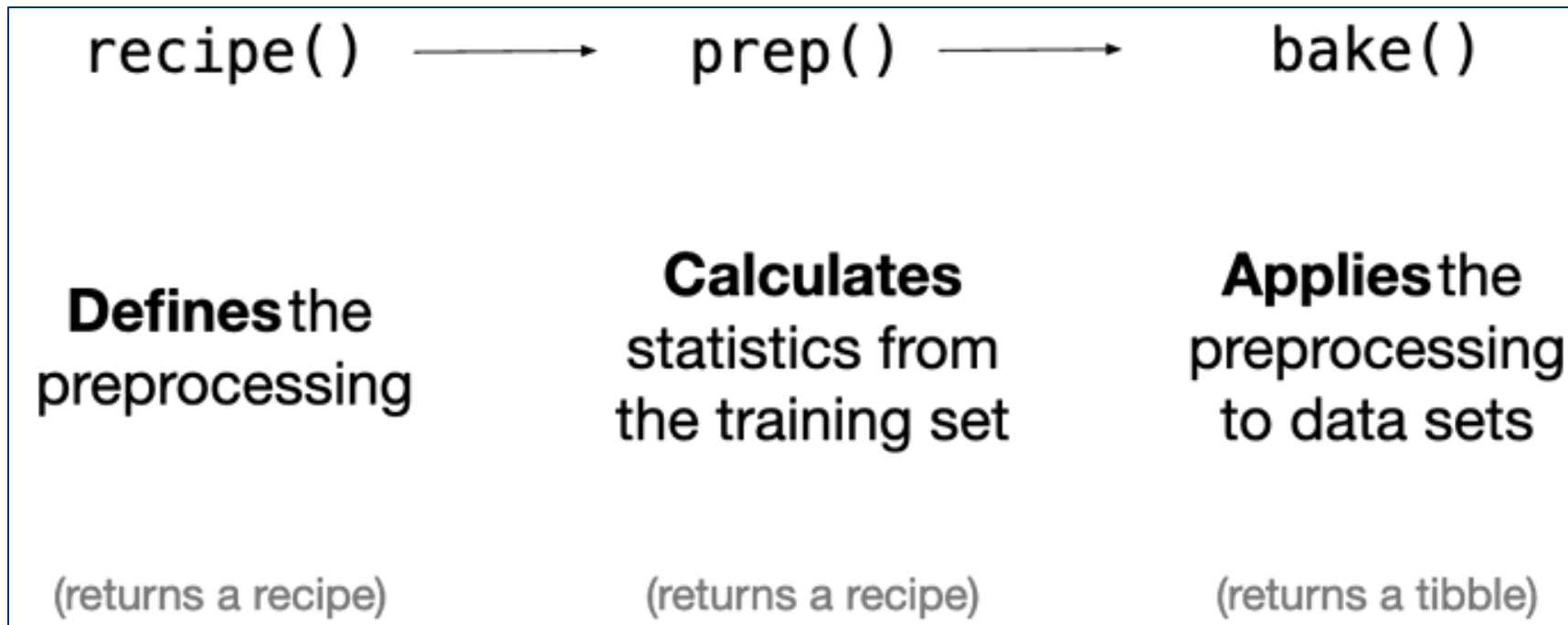
Some commonly used cut-offs include:

- 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()`

The background of the slide is a light gray color, decorated with a pattern of small, semi-transparent hexagons and dots in various colors including blue, yellow, green, orange, and white. These shapes are scattered across the entire page, creating a festive or celebratory feel.

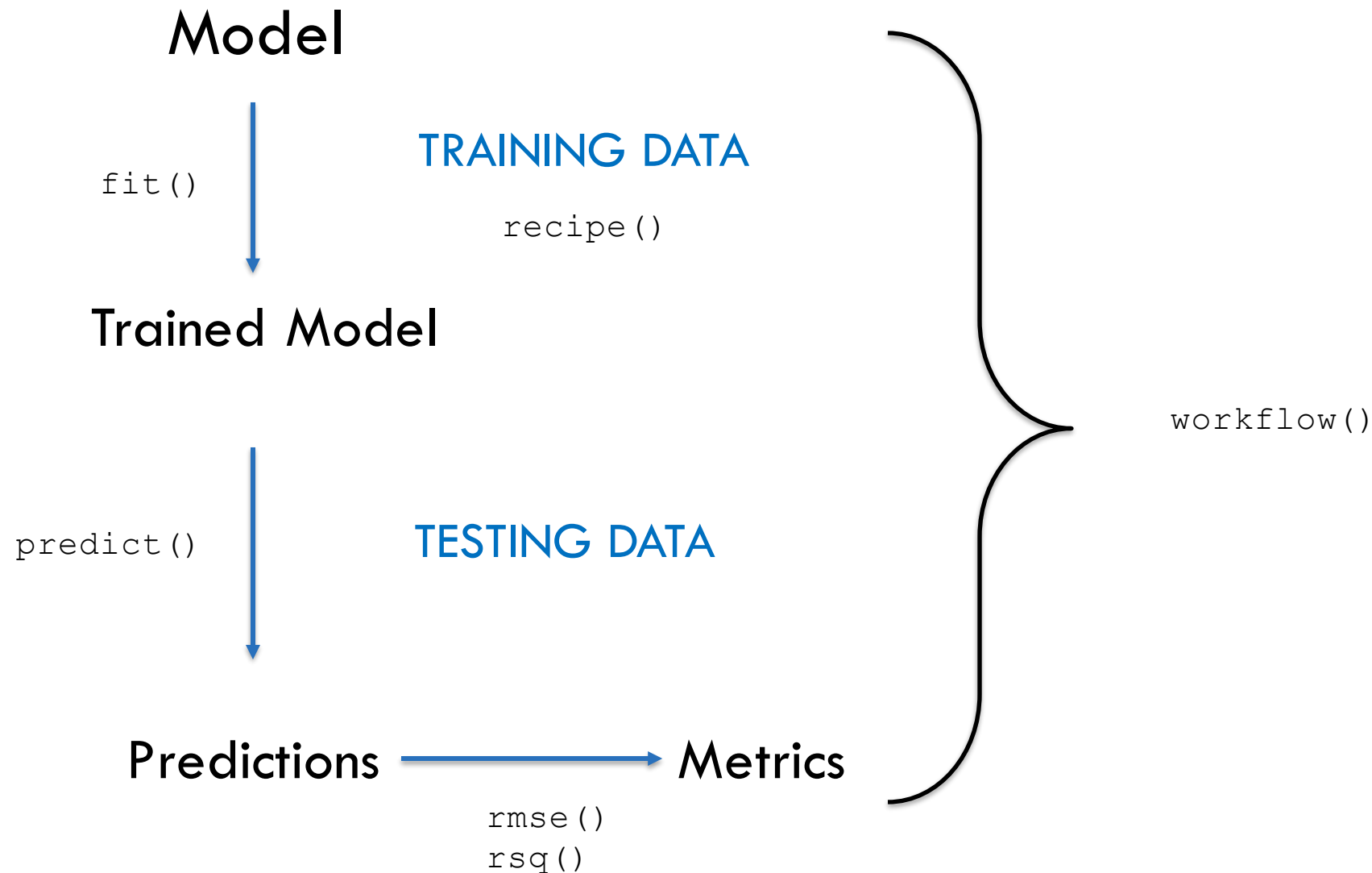
Let's make our life easier!

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()`)

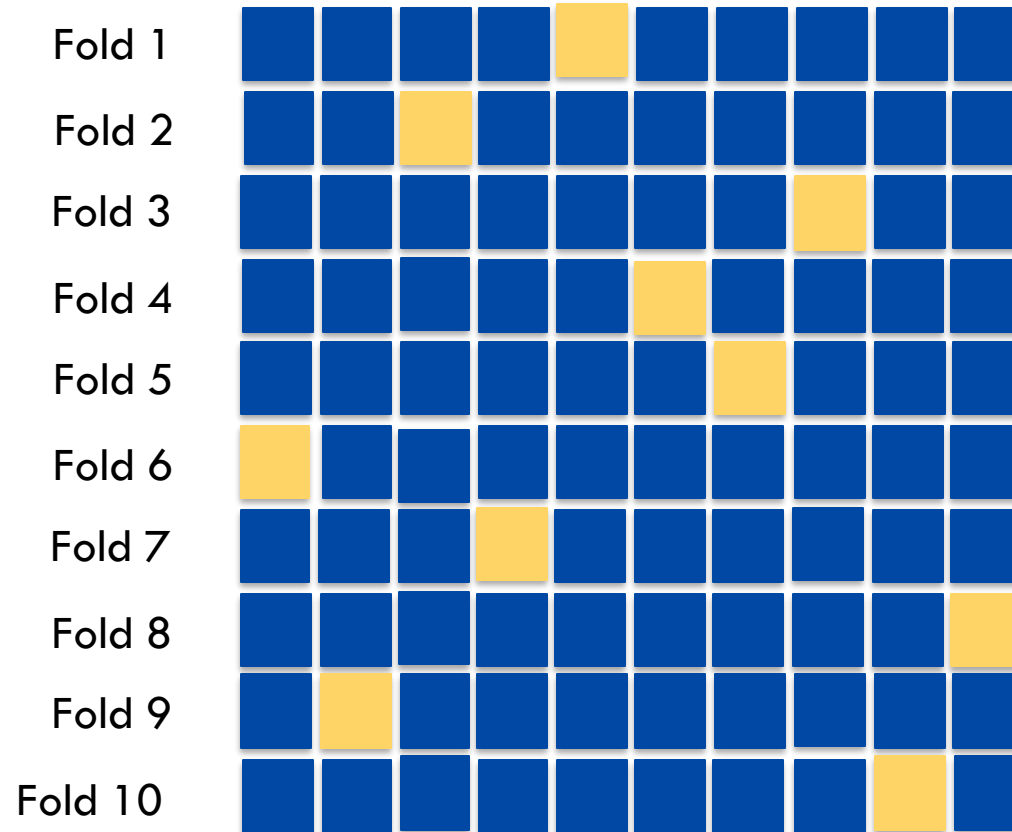
Within `predict()`, values from the **training** data are used to standardize the **testing** data.



Resampling – cross-validation

Pre-Process

Original
Training
Testing



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.feats.engineering/>
- <https://smltar.com/>



Your feedback is important!