RStudio®

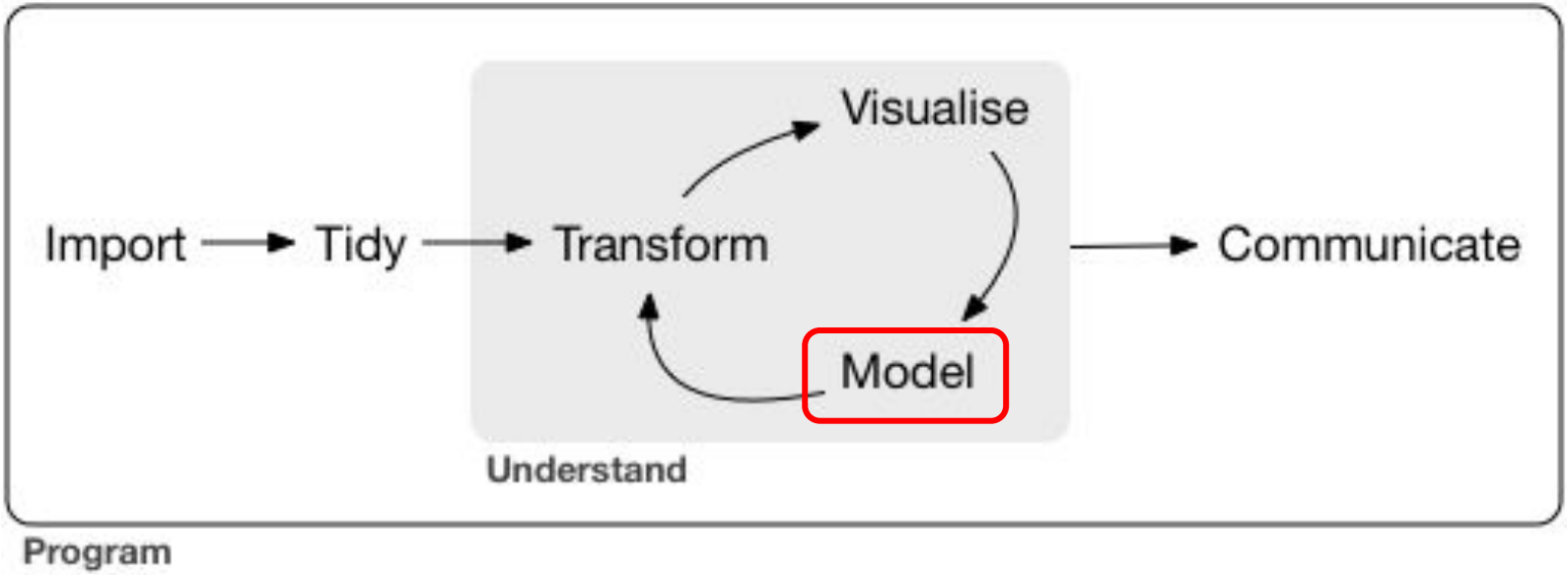
Machine Learning in R with `tidymodels`

Nick Rohrbaugh
RStudio Customer Success
nick@rstudio.com

github.com/nrohr/learn-tidymodels

Why model?

A typical data science project



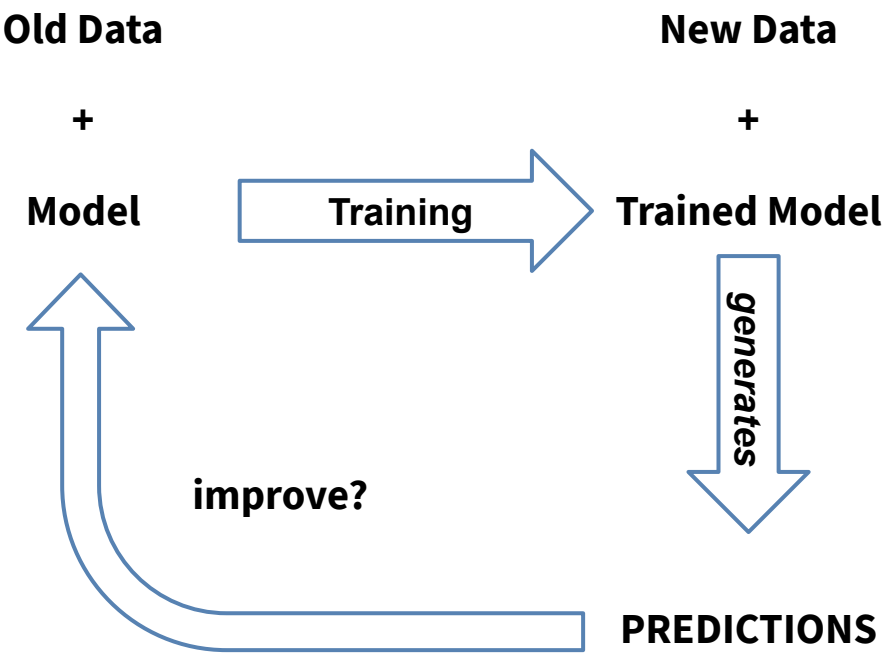
From *R for Data Science*

Why model?

1. To understand our data (and our world)
2. To **predict** things

The goal of machine learning is to **construct models** that **generate accurate predictions** for **future, yet-to-be-seen data**.

Machine Learning



Why model **in R**?

Why model in R?

- R has cutting edge models
- It's easy to integrate your work in R with other tools and languages: C, C++, tensorflow, keras, python, Spark, stan, …
- I like R's tools for data wrangling and visualization/communication

Why **not** model in R?

Modeling in R has a few downsides…

- R was (is) built for ease of use, not performance. (It's not C)
- R almost always requires data to be in memory (minus a few exceptions)
- The process was *inconsistent* across methods/packages
    - There are several methods for specifying terms in a model; not all packages support all methods
    - 99% of model functions auto-generate dummy variables (but 1% don't)
    - Sparse matrices can be used (unless they can't)

tidymodels

* and some more

RStudio

What is `tidymodels`?

- An R package for modeling and machine learning using tidyverse principles
- A "package of packages" (like `tidyverse`) with ~9 core packages
- A framework for consistent, organized machine learning processes in R

Tutorials, articles, and more resources: tidymodels.org

Tidy Modeling with R book by Max Kuhn and Julia Silge: tmwr.org

Modeling with `tidymodels`

Pre-processing → **Train model** → **Validate & select model**

RStudio

Our data

```
library(tidyverse)
library(tidymodels)

data(credit_data)
```

(Simulated) credit data from modeldata package, part of tidymodels

Can we predict Status (default)?

Status <fctr>	Seniority <int>	Home <fctr>	Time <int>	Age <int>	Marital <fctr>	Records <fctr>	Job <fctr>	Expenses <int>	Income <int>	Assets <int>	Debt <int>	Amount <int>	Price <int>
good	9	rent	60	30	married	no	freelance	73	129	0	0	800	846
good	17	rent	60	58	widow	no	fixed	48	131	0	0	1000	1658
bad	10	owner	36	46	married	yes	freelance	90	200	3000	0	2000	2985
good	0	rent	60	24	single	no	fixed	63	182	2500	0	900	1325
good	0	rent	36	26	single	no	fixed	46	107	0	0	310	910
good	1	owner	60	36	married	no	fixed	75	214	3500	0	650	1645
good	29	owner	60	44	married	no	fixed	75	125	10000	0	1600	1800
good	9	parents	12	27	single	no	fixed	35	80	0	0	200	1093
good	0	owner	60	32	married	no	freelance	90	107	15000	0	1200	1957
bad	0	parents	48	41	married	no	partime	90	80	0	0	1200	1468
good	6	owner	48	34	married	no	freelance	60	125	4000	0	1150	1577
good	7	owner	36	29	married	no	fixed	60	121	3000	0	650	915
good	8	owner	60	30	married	no	fixed	75	199	5000	2500	1500	1650
good	19	priv	36	37	married	no	fixed	75	170	3500	260	600	940
bad	0	other	18	21	single	yes	partime	35	50	0	0	400	500

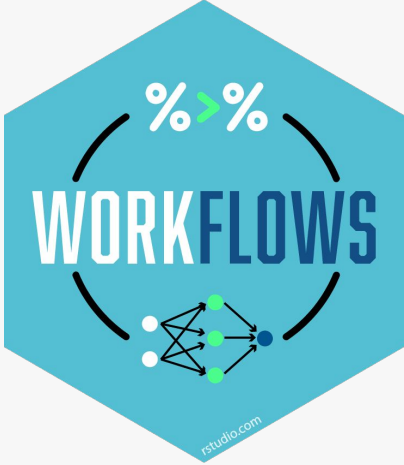
1–15 of 4,454 rows

Modeling with `tidymodels`

1. Split our data into training and testing sets with **`rsample`**
2. Do additional preprocessing and feature engineering with **`recipes`**
3. Specify our model with **`parsnip`** and fit it to training data
4. Use **`workflows`** to preprocess training and testing data separately, to avoid data leakage

RStudio

Workflows



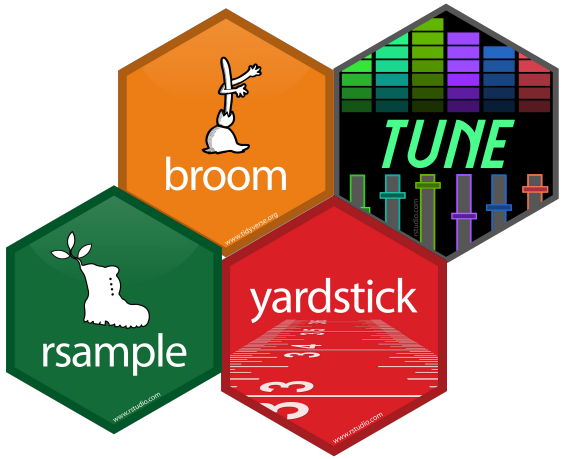
Workflows pair recipes with models. Workflows will help us:

1. Process a recipe with our training data (calculate means & SDs, which vars to remove, etc.)
2. Apply the recipe to our training data
3. Apply the same recipe to our test data (without recomputing anything from step 1)

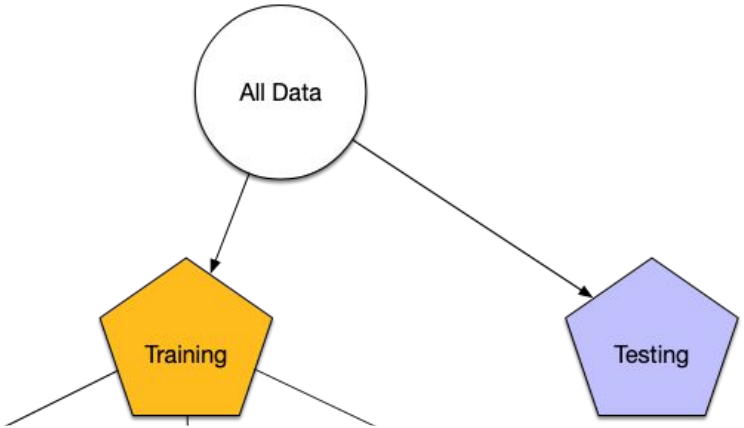
without having to keep close track of separate objects in our workspace.

Modeling with `tidymodels`

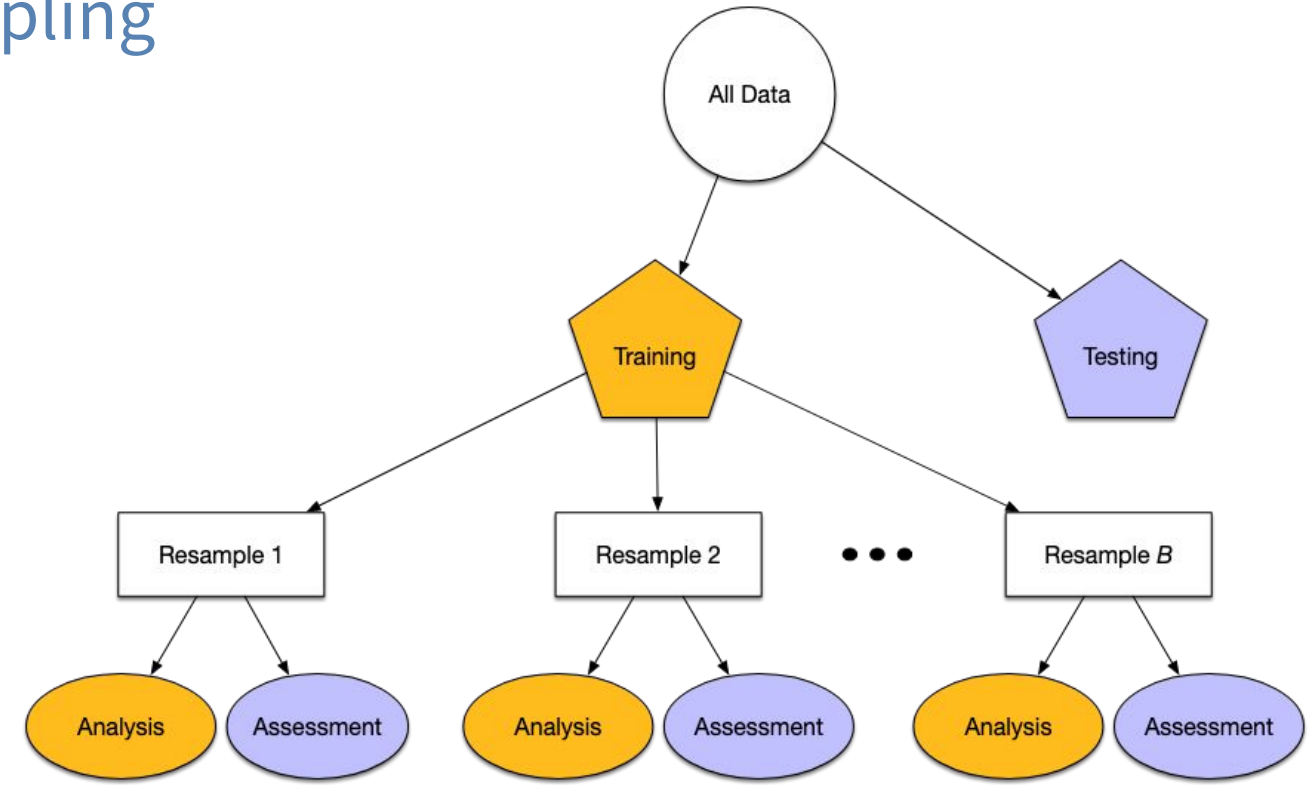
5. Generate predictions against test data, then use **`broom`**, **`yardstick`**, and **`tune`** to check fit.
6. Evaluate model with resampling with **`rsample`**



RStudio



Resampling



Kuhn & Johnson (2019)

RStudio

Additional resources

- Tidymodels website: tidymodels.org
- Tidy Modeling with R book: tmwr.org
- Julia Silge's blog, YouTube channel, and online course
- Alison Hill's workshop materials (rstudio::conf 2020)

thanks!

h/t Alison Hill, Max Kuhn, Julia Silge, Tom Mock, Garrett Grolemund & others @

RStudio

Nick Rohrbaugh
nick@rstudio.com