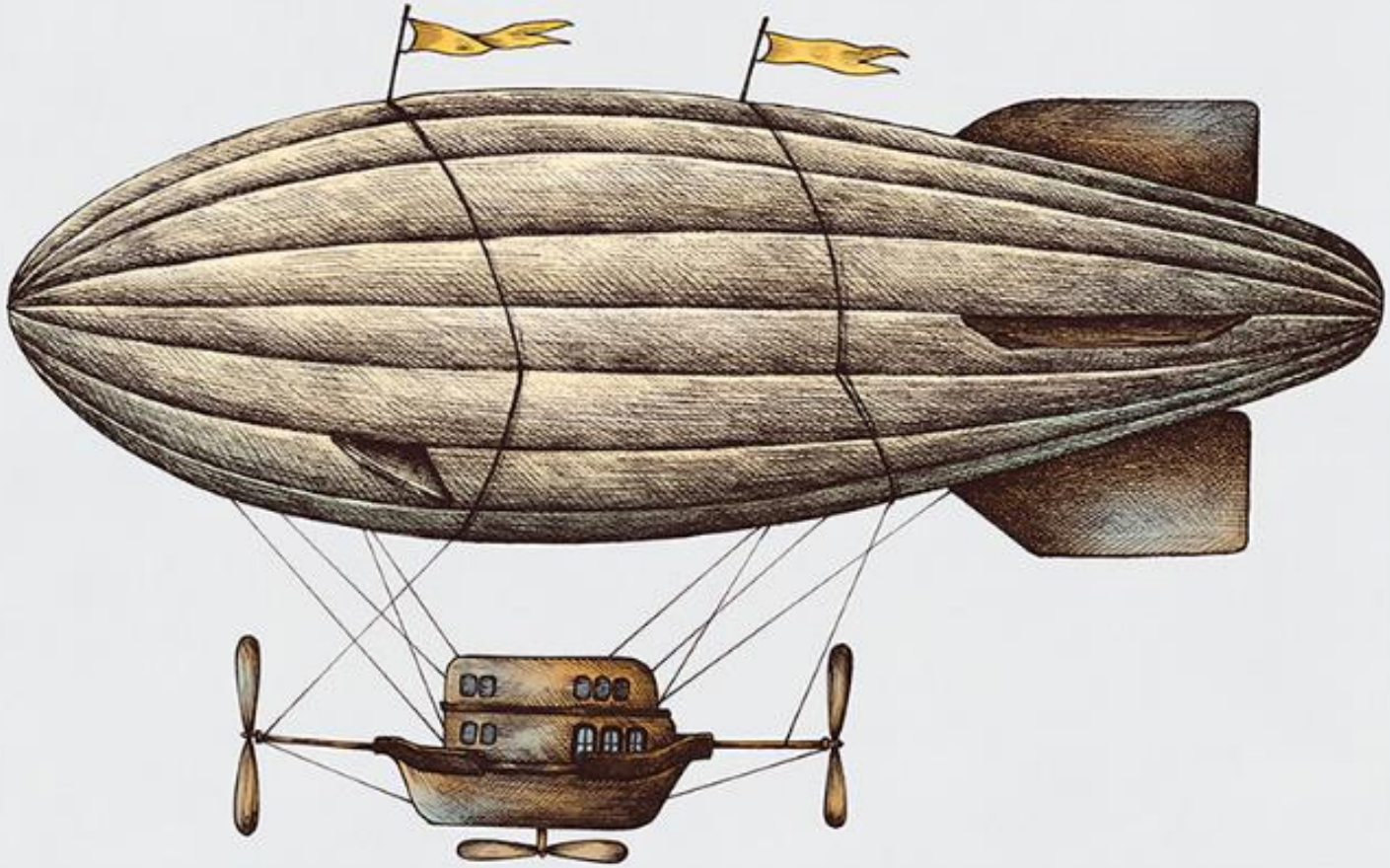
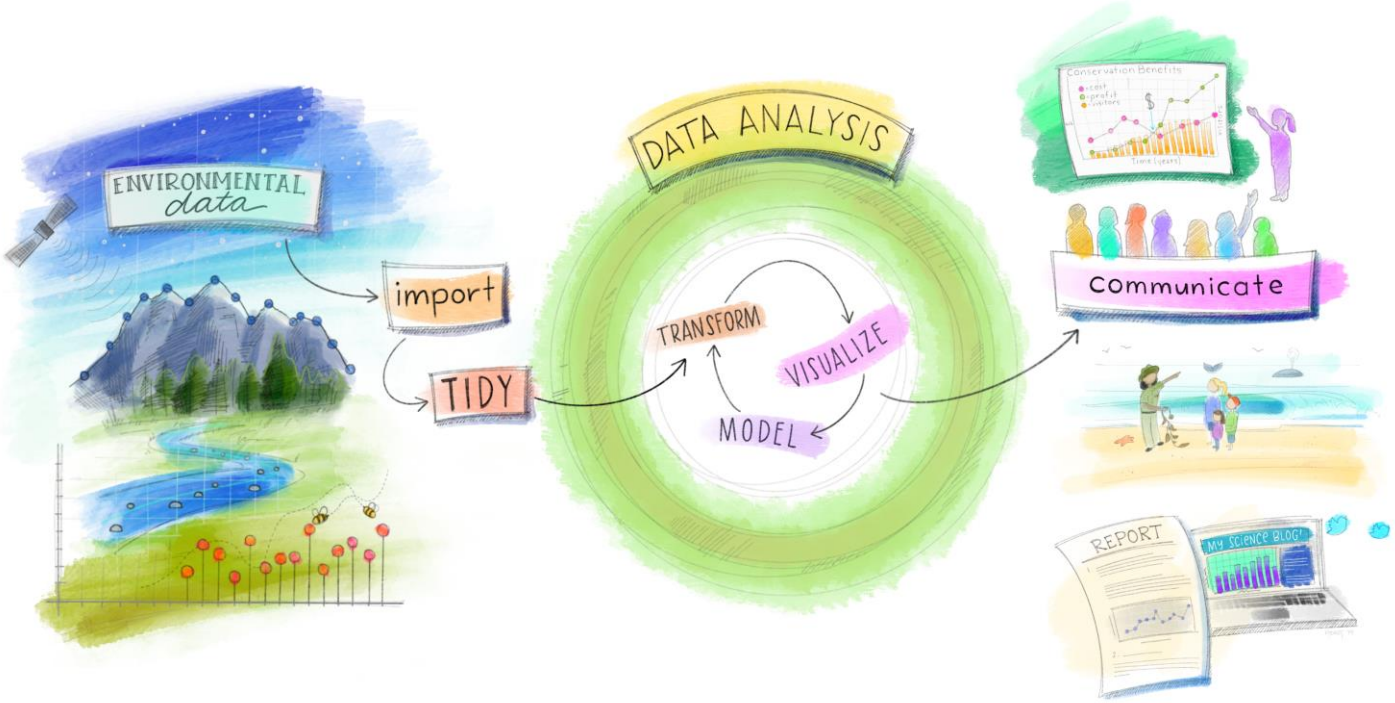
# A quick flight to the edge of Data Science



Eric Wanjau
Gold MLSA
Researcher/Data Scientist LIDA

## Modeling and Machine Learning with Tidymodels

# How does modelling fit into the Data Analysis process



There are always a few critical phases of data analysis that come before modeling.

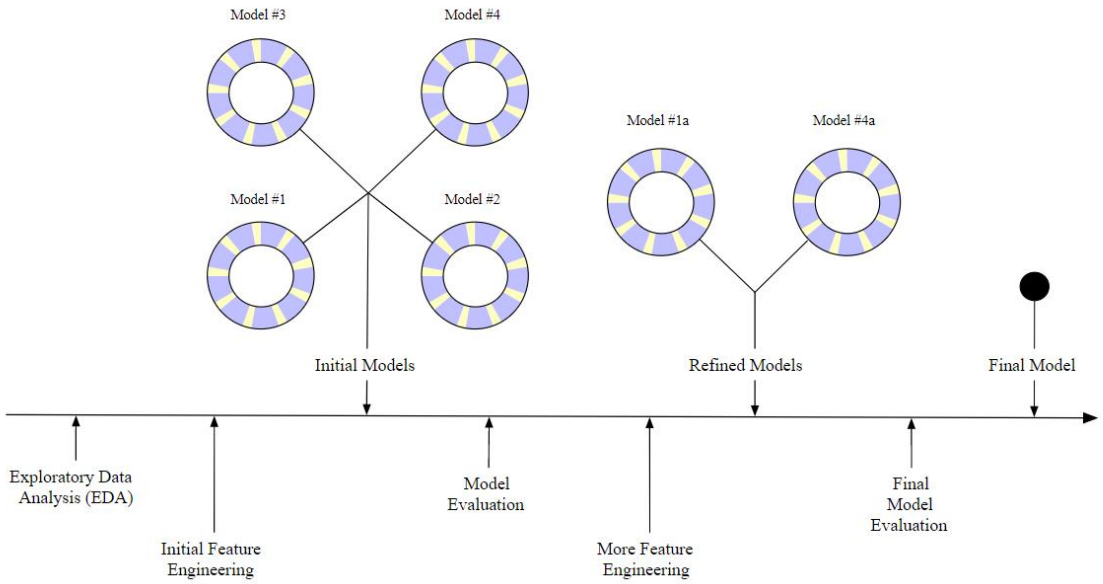
# The modeling process

- Many books/courses often portray predictive modeling as a one-time easy peasy process.

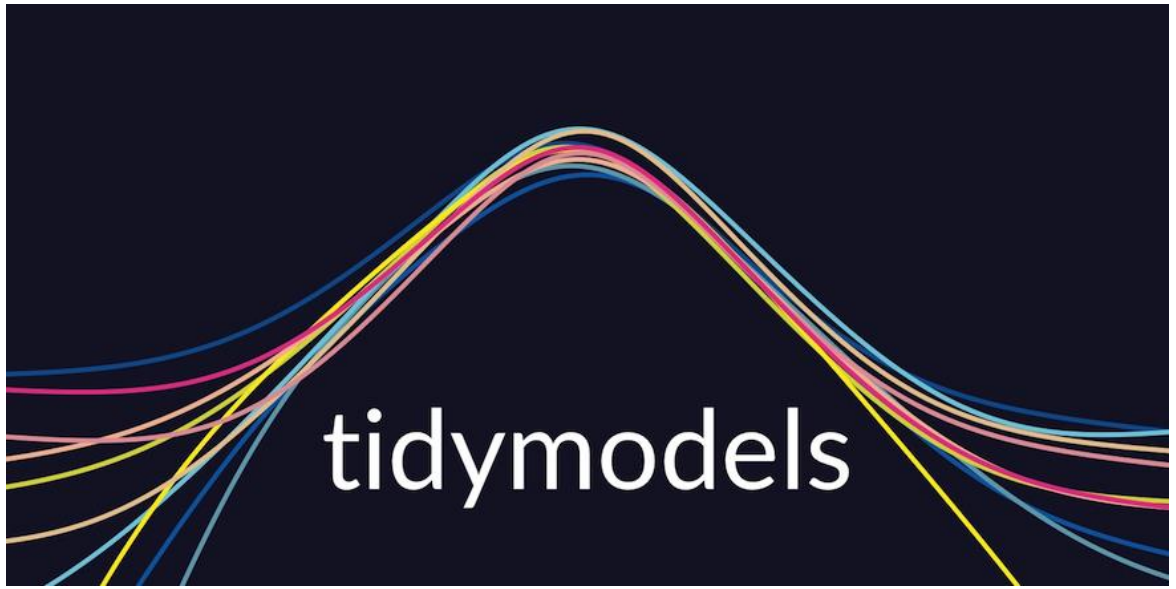
- A better analogy would be to view it as an iterative process involving a number of steps and decisions to make.

# The modeling process

- Exploratory data analysis (EDA)
- Feature engineering
- Model tuning and selection (large circles with alternating segments)
- Moving on to the final model



There are many modeling packages in R and then there is




Modeling framework that allows you to build upon existing R packages

# Why Tidymodels?

# 1. We are all on the same team

unified interface to underlying packages that do the same thing

```r
# From ranger
rf_2 <- ranger(
  y ~ .,
  data = dat,
  mtry = 10,
  num.trees = 2000,
  importance = "impurity"
)
```

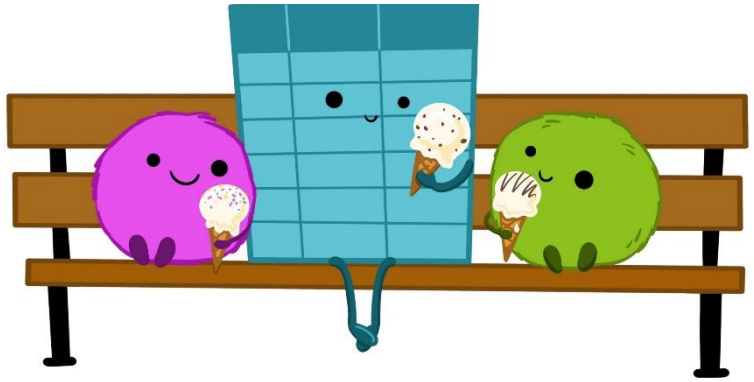
```r
# From randomForest
rf_1 <- randomForest(
  y ~ .,
  data = .,
  mtry = 10,
  ntree = 2000,
  importance = TRUE
)
```

```r
# From sparklyr
rf_3 <- ml_random_forest(
  dat,
  intercept = FALSE,
  response = "y",
  features = names(dat)[names(dat) != "y"],
  col.sample.rate = 10,
  num.trees = 2000
)
```

# Model spec in tidymodels

```
rf <- rand_forest(mtry = 10, trees = 2000) %>%
        set_mode("regression") %>%
        set_engine("_ _ _ _ _ _")
```

# 2. Returns our good old tibbles



```
## # A tibble: 6 x 4
##       a     b     c lambda
##   <dbl> <dbl> <dbl>  <dbl>
## 1 0.333 0.333 0.333   1
## 2 0.333 0.333 0.333   1
## 3 0.333 0.333 0.333   0.1
## 4 0.333 0.333 0.333   0.1
## 5 0.373 0.244 0.383   0.01
## 6 0.327 0.339 0.334   0.01
```

```
predict(three_class_mod, newx = new_x,
        type = "response")
```

```
## , , s0
##
##                a     b     c
## sample_1 0.333 0.333 0.333
## sample_2 0.333 0.333 0.333
##
## , , s1
##
##                a     b     c
## sample_1 0.333 0.333 0.333
## sample_2 0.333 0.333 0.333
##
## , , s2
##
##                a     b     c
## sample_1 0.373 0.244 0.383
## sample_2 0.327 0.339 0.334
```

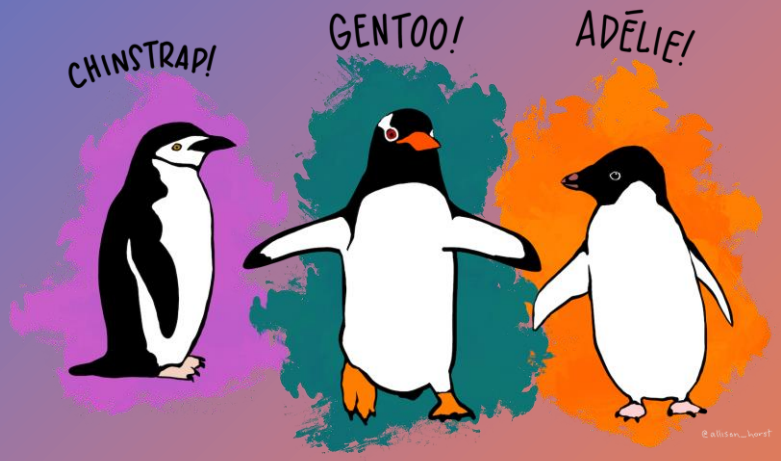
# 3. Additional tools for creating effective and high-quality models


Splitting data

Feature Engineering

Model performance

# Classifying penguins.




- Build and fit a random forest model
- Use the model to classify
- Try out a different engine

# Knowledge Check

You plan to use the Tidymodels framework to train a model. What information would you need to make a model specification? (there are multiple answers)

A. Model Engine
B. Model Mode
C. Model Recipe
D. Model Type

# Knowledge Check

You plan to use the Tidymodels framework to train a model. What information would you need to make a model specification?

A. Model Engine: is the R package which will be used to fit the model

B. Model Mode: includes common options like regression and classification

C. Model Recipe: A recipe is an object that defines a series of steps for data processing.

D. Model Type:  differentiates models such as logistic regression, decision tree models

# Q. What's a penguin's favorite movie?

Frozen

# Recap & Further Reading

- ✓ What is the Tidymodels framework?
- ✓ How Tidymodels eases modeling for practitioners from diverse backgrounds.
- ✓ Building models using Tidymodels framework.

- Max Kuhn and Julia Silge, *Tidy Modeling with R*.
- *Max Kuhn talk – Cleveland RUG*
- Tidymodels reference website.
- Bradley Boehmke & Brandon Greenwell, *Hands-On Machine Learning with R*.
- H. Wickham and G. Grolemund, *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*.