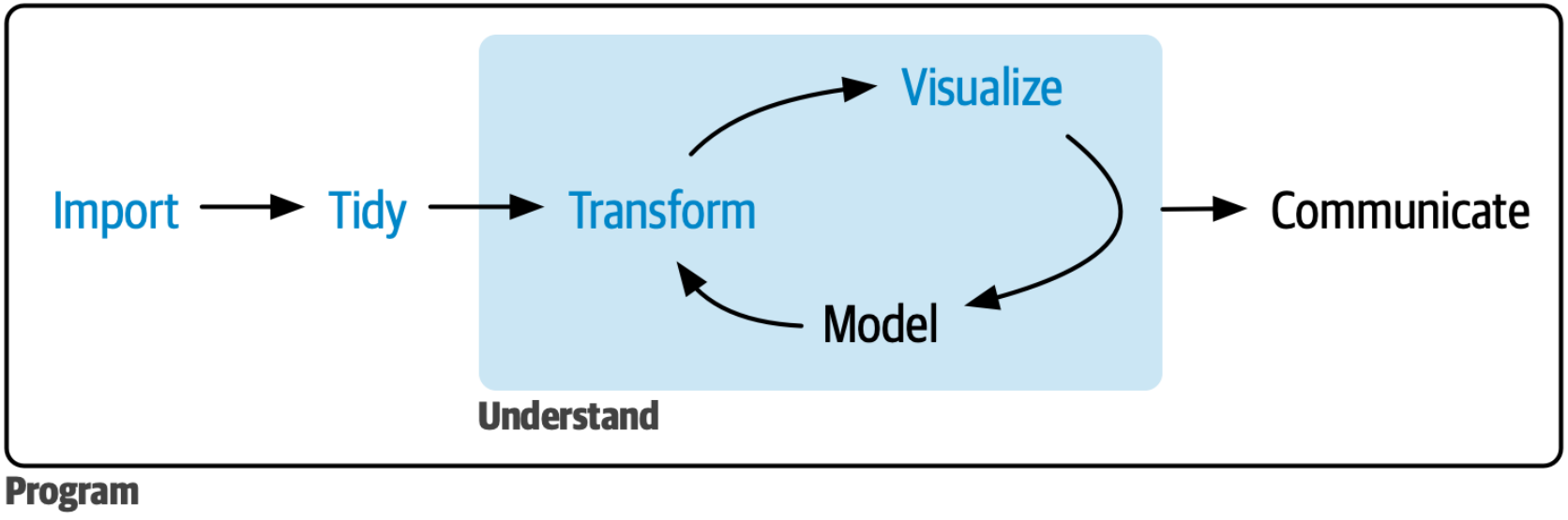
https://r4ds.hadley.nz/ Chapter 3-9

CSCI 297b, Spring 2023

April 30, 2023

The Big Picture



The dplyr package and the nycflights13 dataset

```
library(nycflights13)
library(tidyverse)
```

the nycflights13 dataset

```
> glimpse(flights)
Rows: 336,776
Columns: 19
$ year           <int> 2013, 2013, 2013, 2013,...
$ month          <int> 1, 1, 1, 1, 1, 1, 1, 1,...
$ day            <int> 1, 1, 1, 1, 1, 1, 1, 1,...
$ dep_time       <int> 517, 533, 542, 544, 554...
$ sched_dep_time <int> 515, 529, 540, 545, 600...
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5...
$ arr_time       <int> 830, 850, 923, 1004, 81...
$ sched_arr_time <int> 819, 830, 850, 1022, 83...
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 1...
$ carrier        <chr> "UA", "UA", "AA", "B6",...
$ flight         <int> 1545, 1714, 1141, 725, ...
$ tailnum        <chr> "N14228", "N24211", "N6...
$ origin         <chr> "EWR", "LGA", "JFK", "J...
$ dest           <chr> "IAH", "IAH", "MIA", "B...
$ air_time       <dbl> 227, 227, 160, 183, 116...
$ distance       <dbl> 1400, 1416, 1089, 1576,...
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6,...
$ minute         <dbl> 15, 29, 40, 45, 0, 58, ...
$ time_hour      <dttm> 2013-01-01 05:00:00, 2...
```

The dplyr package

- The first argument is always a data frame.
- The subsequent arguments typically describe which columns to operate on, using the variable names (without quotes).
- The output is always a new data frame.
- Each verb operates on either
  - rows,
  - columns,
  - groups, or
  - tables

The pipe

```
          x |> f(y)   ⇔   f(x, y)
x |> f(y) |> g(z)   ⇔   g(f(x, y), z)
```

```
flights |>
  filter(dest == "IAH") |>
  group_by(year, month, day) |>
  summarize(
    arr_delay = mean(arr_delay, na.rm = TRUE)
  )
```