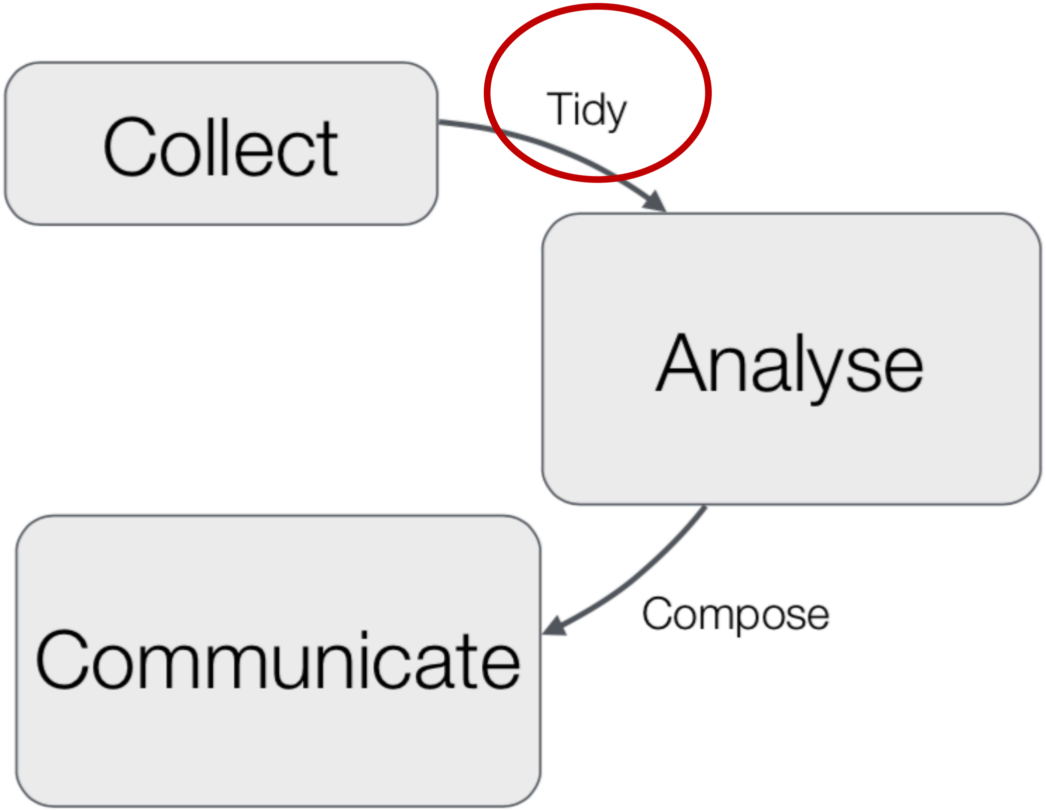


Key concepts:

Workflows, Tidy Data, Relational Data

What is a workflow?

A series of steps necessary to complete a task

For example, I hate flight delays! I want to avoid days and times that have a lot of delays.

To minimize my risk of flight delays, I should take a large dataset of flights,

then select those flights that had delays,

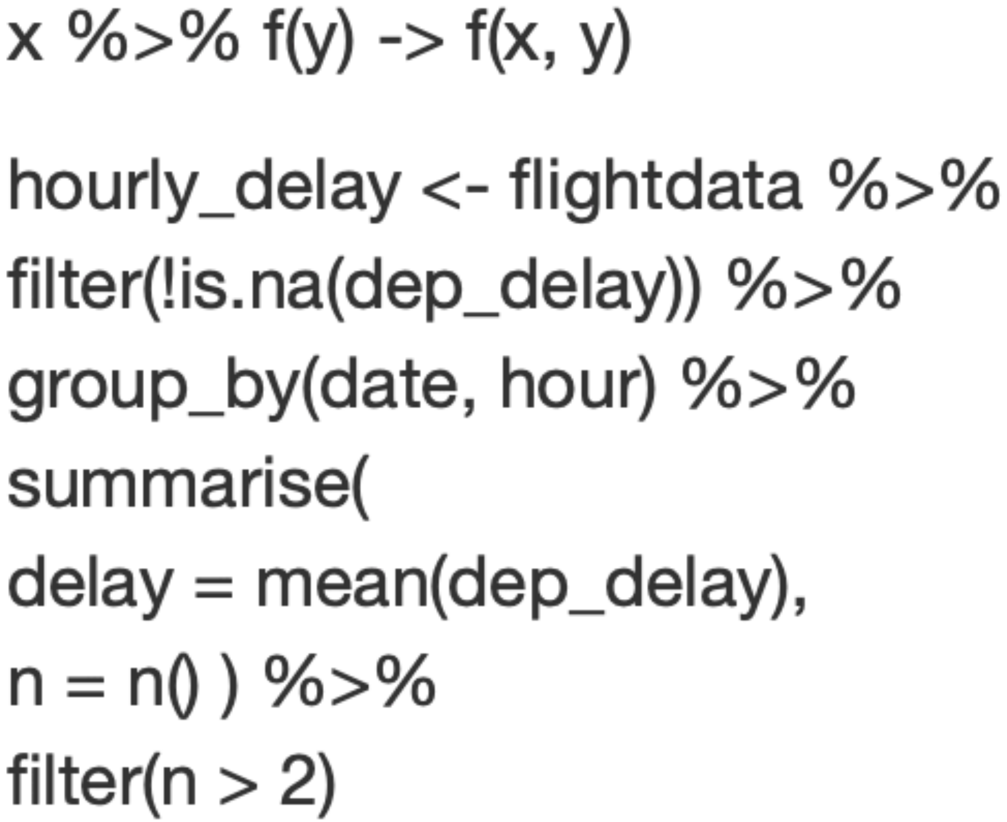
then group those flights by date and hour,

then summarize by mean flight delay,

then select those times with delays over 2 hours

Then avoid traveling on those days and times!

dplyr allows me to do just that with the “pipeline”



```
x %>% f(y) -> f(x, y)

hourly_delay <- flightdata %>%
filter(!is.na(dep_delay)) %>%
group_by(date, hour) %>%
summarise(
delay = mean(dep_delay),
n = n() ) %>%
filter(n > 2)
```

dplyr

- **filter:** keep rows that match a criteria
- **select**: pick columns by name
- **arrange:** reorder rows
- **mutate:** add new variables
- **summarize:** reduce variables to values
- **+ group_by**