WRITING WITH DATA

Jeff Goldsmith, PhD

Department of Biostatistics

Writing is important

- You're going to spend a lot of your time communicating in writing
  - With collaborators, a general public, future you
  - About data cleaning, analyses, results
  - In formal reports, brief summaries, replies to questions

- Time to get good

Tools

- Code is necessary but not sufficient
- Use tools that combine your code and text
- Greatly facilitates reproducibility, which is a big concept
  - In short, someone you don't know or work with should be able to reproduce each step of your analysis
  - As a part of this, they should understand why you did what you did
  - (Again, this someone is often future you)

- We'll use R Markdown to write reproducible reports

General tips

- Know your audience
  - Are they statistically knowledgeable?
  - How many details do they want / need?

- Say exactly what you did
  - Don't leave any thing important out
  - Not the same as a step-by-step list of what you typed into R

General structure

- Introduction / overview
- Data and methods
  - File names
  - Summary statistics
  - Exploratory analysis
  - Formal analysis
- Results
- Discussion

- Some version of these exist in almost everything I write
- Sometimes these are long, sometimes they're a sentence

Introduction

- What is the context for this problem?
- What kind of data were gathered?
- What do you hope to learn?

Data

- Importing, tidying, and editing
  - Loading data
  - Reorganizing into usable form
  - Identifying missing values
  - Recoding and creating variables
- Summary statistics
  - Sample size
  - Means or proportions of major variables

Methods / "models"

- Exploratory analyses
  – Visualizations
  – Numerical summaries

- Formal analyses
  – Model components
  – Model strategy
  – Formal comparisons of interest, tests, significance levels

Results

- What did you find in exploratory analyses (any missing values? data distributions? notable features?)
- What happened in your modeling?
- What is your final model, and what are the important quantities?

Discussion

- What do your results say about the question you hoped to answer?
- What were the limitations of your data or your analysis?
- What open questions remain? Are any of these solvable with the current data?
- What are your next steps?

Some true stuff about writing

- It is not easy
- It takes practice
- It is critical to do well

Recall …
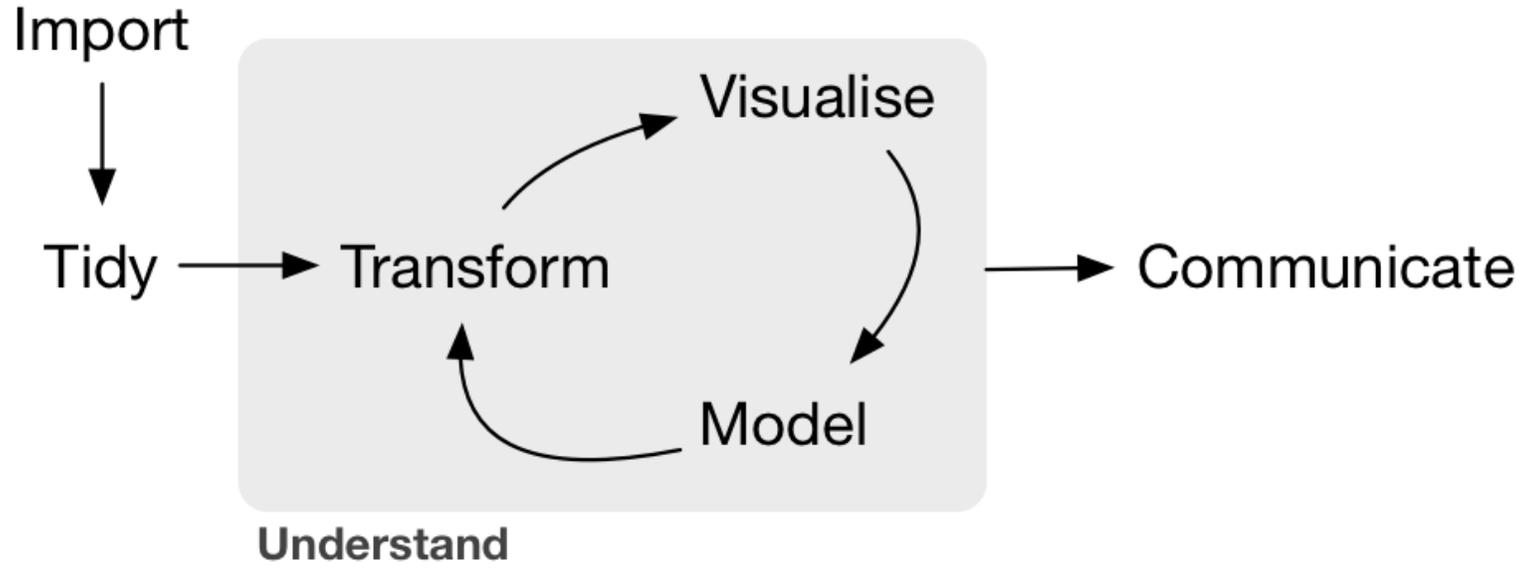
R for Data Science

How analyses are in reality
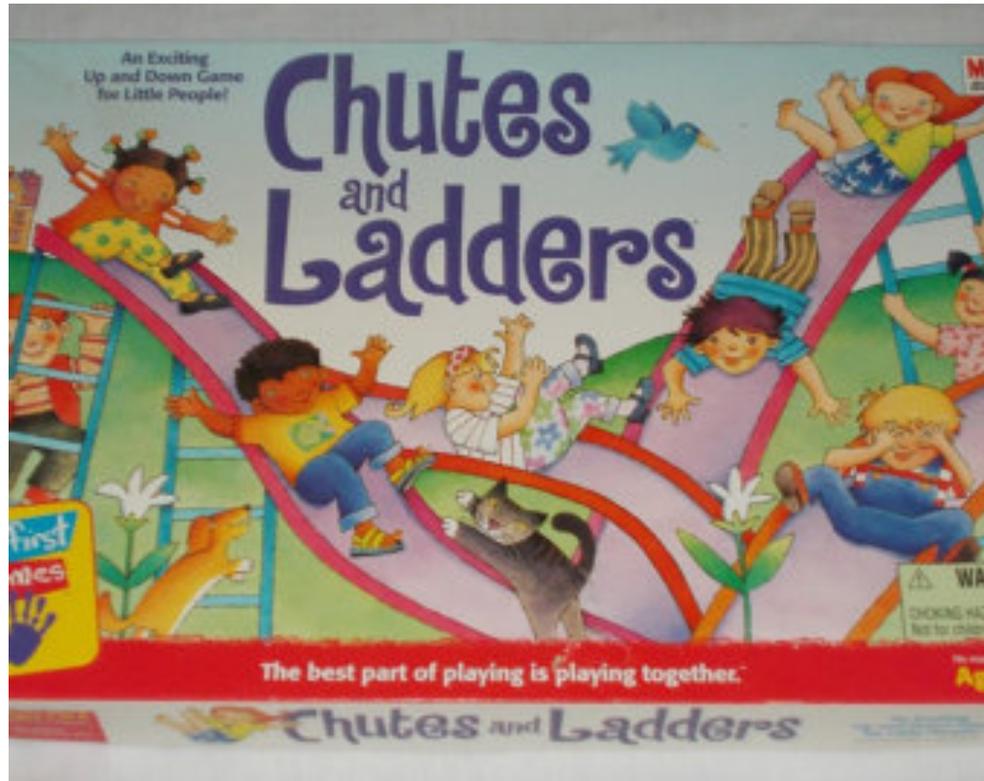
How analyses are in reality
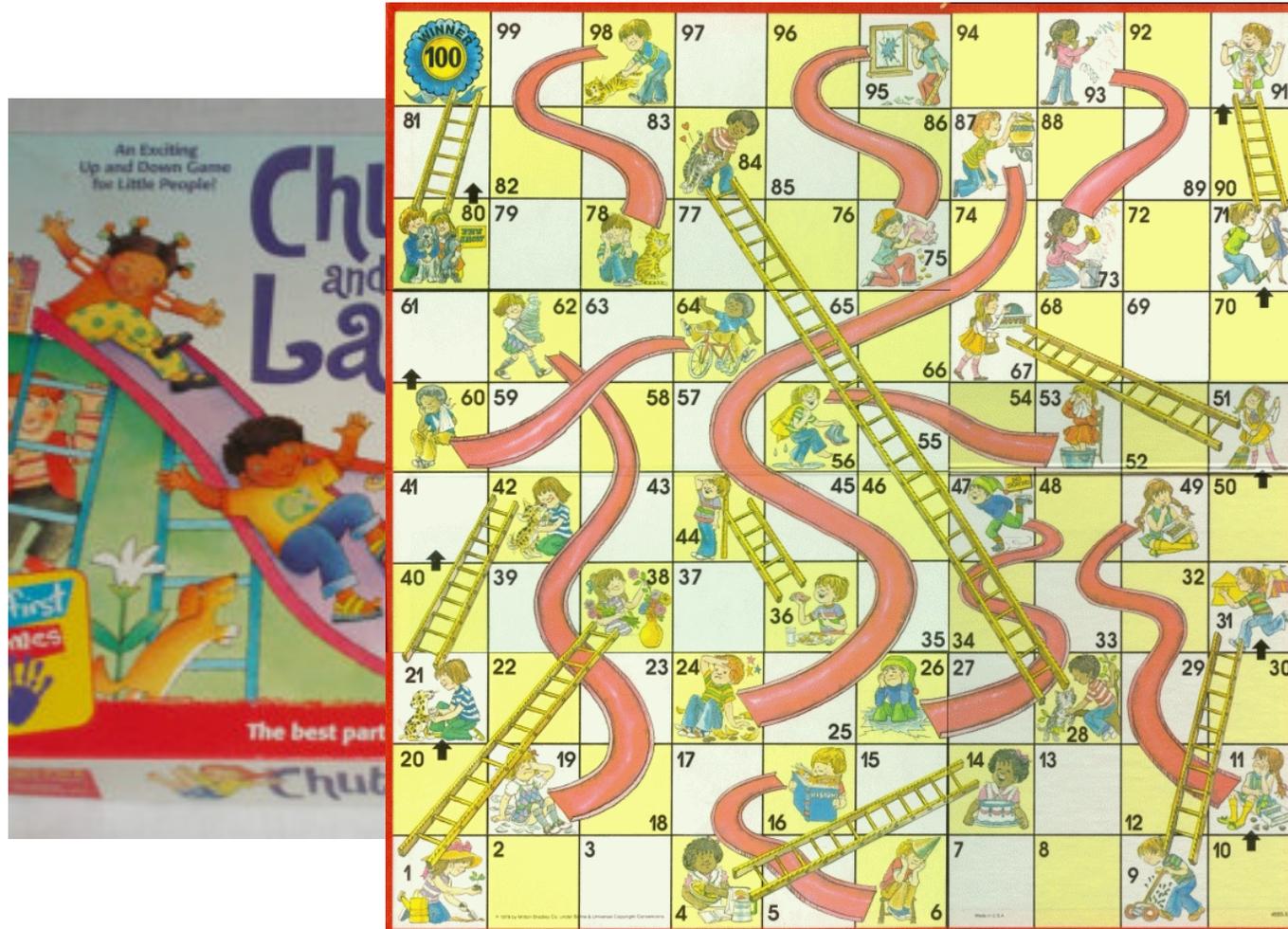
How analyses are presented

Be complete …

… but not too complete.

Be complete …

… but not too complete.

Be complete …

… but not too complete.

Confidence intervals and the fitted value curve in each bootstrap sample is done above.

Part e)

```
#### plot them using ggplot
eachci1<-as.data.frame(eachci)
plot = ggplot() + geom_point(color = "black") +
  geom_line(aes(x=lidar$range, y = UB), color = "blue", linetype = 2) +
  geom_line(aes(x=lidar$range, y = LB), color = "blue", linetype = 2) +
  geom_line(data=eachci1,aes(x=lidar$range, y = V2), color = "green", linetype = 2) +
  geom_line(data=eachci1, aes(x=lidar$range, y = V1), color = "green", linetype = 2)
plot
```

Striking a balance

- This is where practice comes in

R Markdown?

- A "Markdown" language is a lightweight syntax that can be easily converted to HTML or another format (PDF, Word)
- R Markdown lets you combine formatted text with code chunks and the results of those chunks

R for Data Science
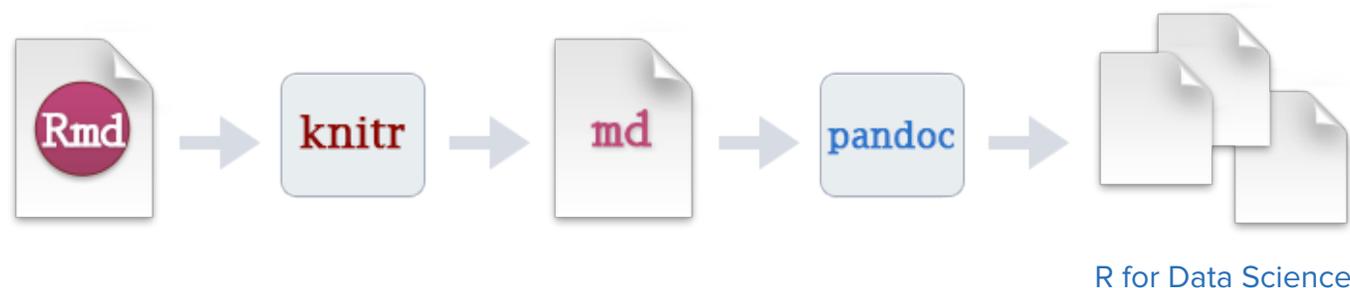
- Having text and code in the same place, and having the combined output be user-friendly, is huge for your workflow

R Markdown?

- A "Markdown" language is a lightweight syntax that can be easily converted to HTML or another format (PDF, Word)
- R Markdown lets you combine formatted text with code chunks and the results of those chunks

R for Data Science

- Having text and code in the same place, and having the combined output be user-friendly, is huge for your workflow