# Introducing R for data analysis

Matthew Brett

# Why R?

- Power
- Range
- Transparency

# Reproducibility crisis

Scientists at Amgen (a drug company) tried to reproduce findings from 53 "landmark" studies.

> *. . . when findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors' direction, occasionally even in the laboratory of the original investigator.*

Of 53 studies, only 6 replicated (11%).

(Begley and Ellis 2012)

# Be transparent in order not to fool yourself

> *The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.*

Richard Feynman, Cargo Cult Science (1974)

# History of R, part 1: S

- The S programming language by John Chambers ; “to turn ideas into software, quickly and faithfully” (Chambers 1998). First version of S in 1976.
- An S program describes the analysis in words (and numbers).
- S is a programming language along with many packages for standard analysis and graphics.

# History of R, part 2: R

- R is a free version of S, developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, first released in 1995.
- Now by far the dominant language in statistics and in many fields of science.
- One of two dominant languages (with Python for data analysis / data science in industry.

# R and other languages

- R and Python
- R and SPSS

# The (ahem) geography of R

- The base R language
- R Studio
- The R Notebook
- The Tidyverse

# Getting help in R

- The R and R studio interface.
- Your search engine.
- Your colleagues.

# The end

Now, to the keyboard.