Reproducible Data Preparation
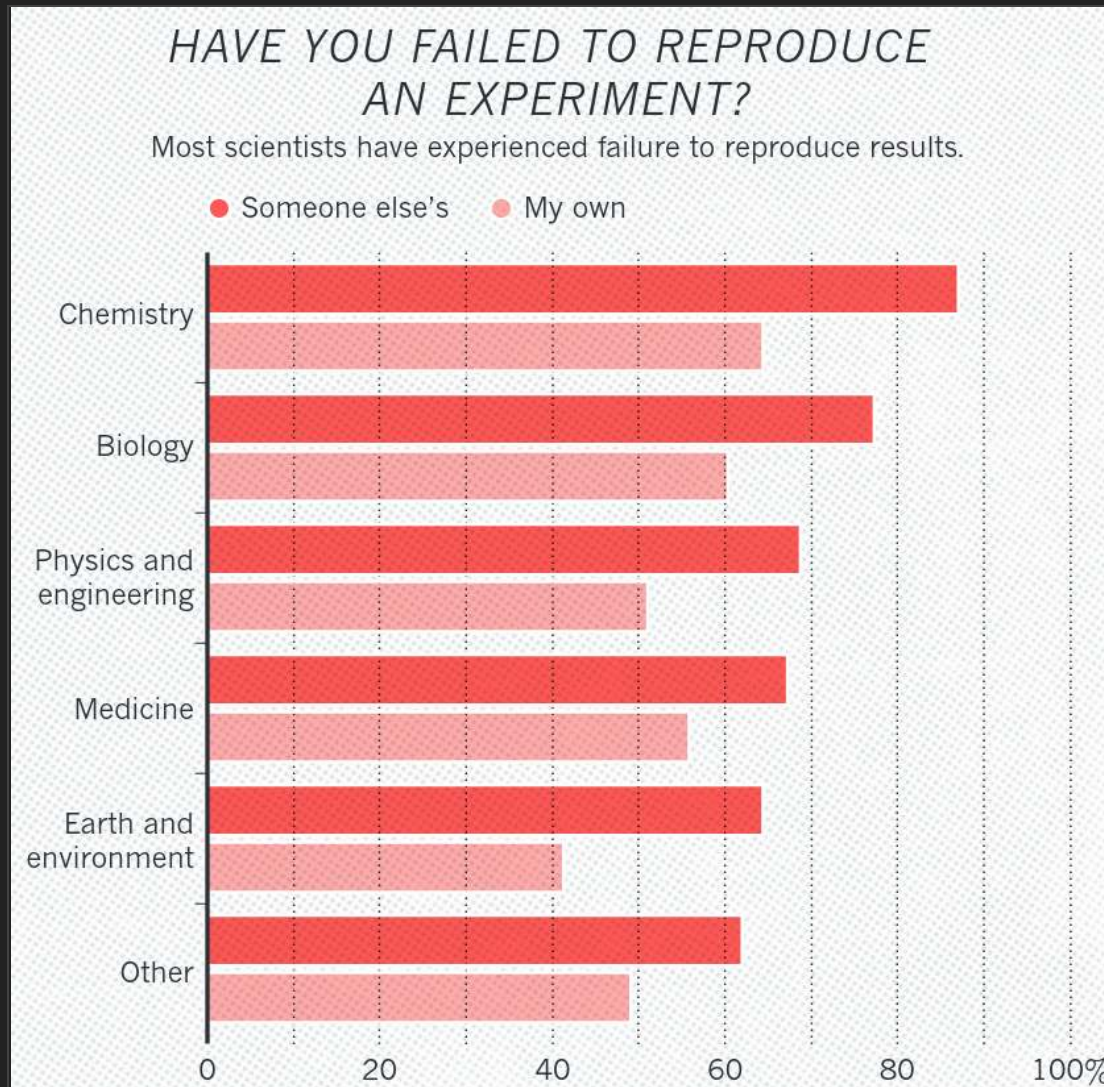
Montreal, CA. Marc-Olivier Jodoin
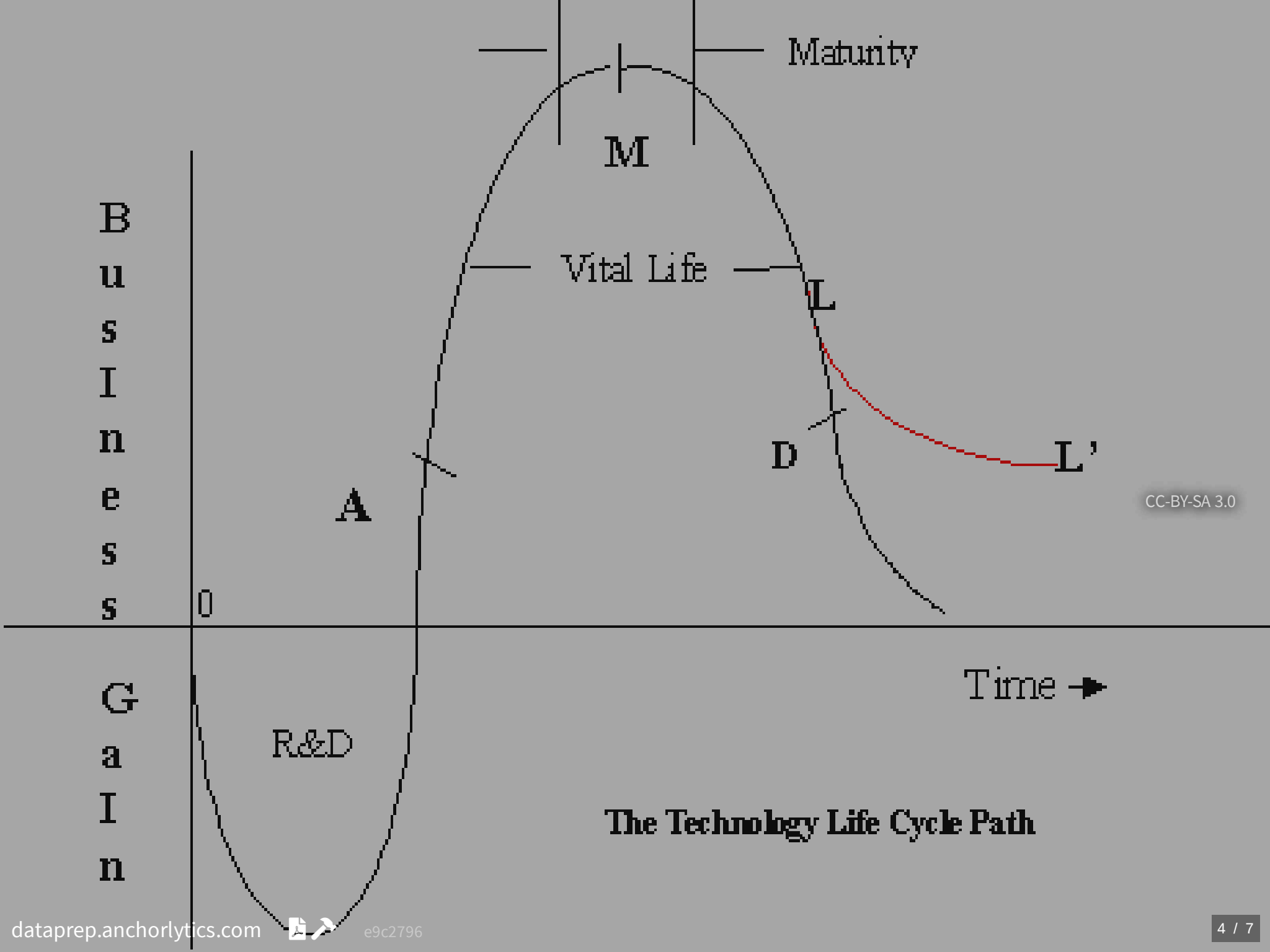
"**Reproducibility** refers to the ability of a researcher to duplicate the **results** of a prior study using the same **materials** as were used by the original investigator.
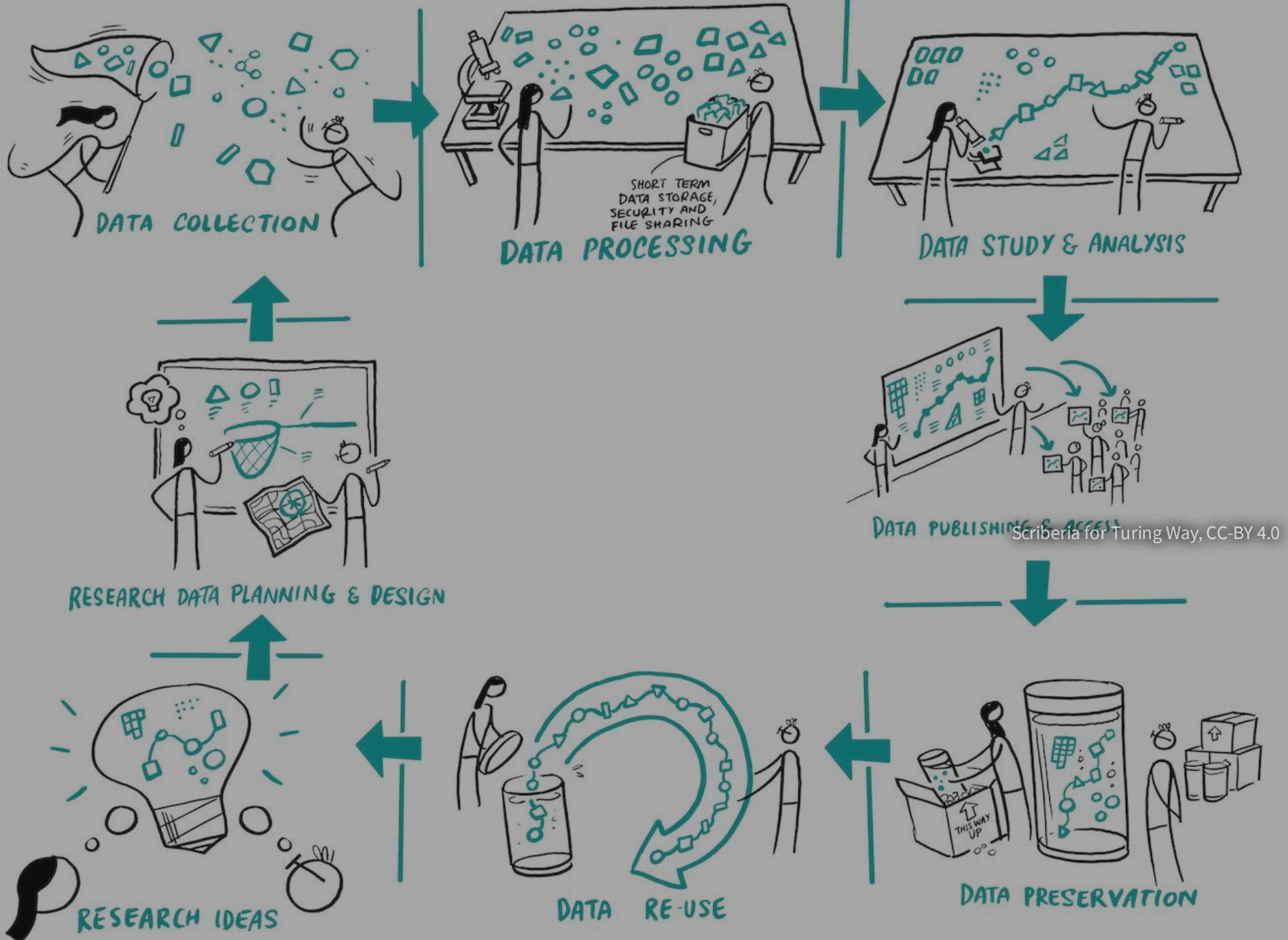
Reproducibility is a **minimum** necessary condition for a finding to be **believable** and **informative**."

Report of the Subcommittee on Replicability in Science Advisory Committee, U.S. National Science Foundation, 2015

e9c2796

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.
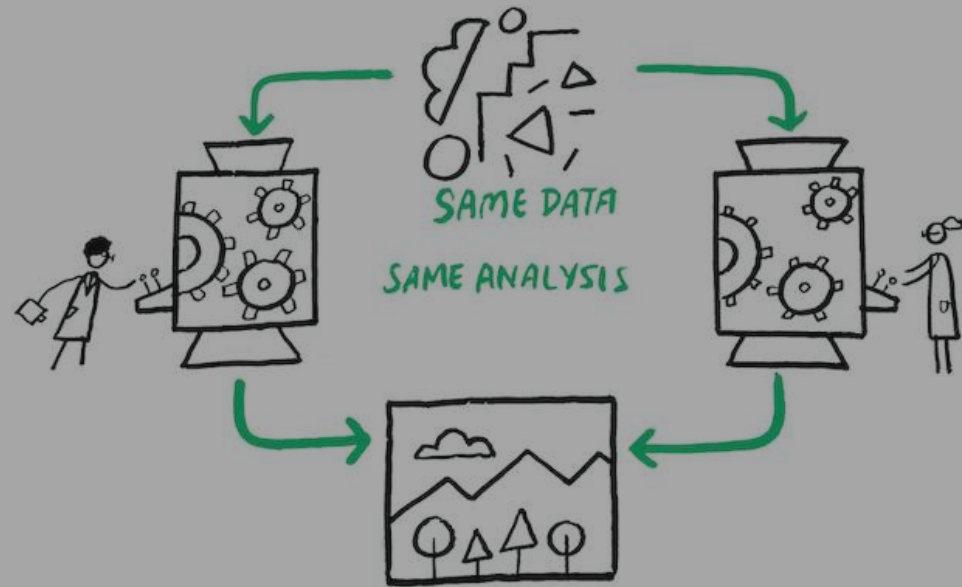
- Psychology: Open Science Collaboration, Science (2015)
- Cancer biology: Begley, C. G. and Ellis, L. M., Nature 483, 531–533 (2012)

The Technology Life Cycle Path

DATA COLLECTION

DATA PROCESSING

SHORT TERM DATA STORAGE, SECURITY AND FILE SHARING

DATA STUDY & ANALYSIS

RESEARCH DATA PLANNING & DESIGN

DATA PUBLISHING & ACCESS

RESEARCH IDEAS

DATA RE-USE

DATA PRESERVATION

THIS WAY UP

Scriberia

REPRODUCIBLE

SAME DATA
SAME ANALYSIS

REPLICABLE

DIFFERENT DATA
SAME ANALYSIS

ROBUST

SAME DATA
DIFFERENT ANALYSIS

GENERALISABLE

DIFFERENT DATA
DIFFERENT ANALYSIS
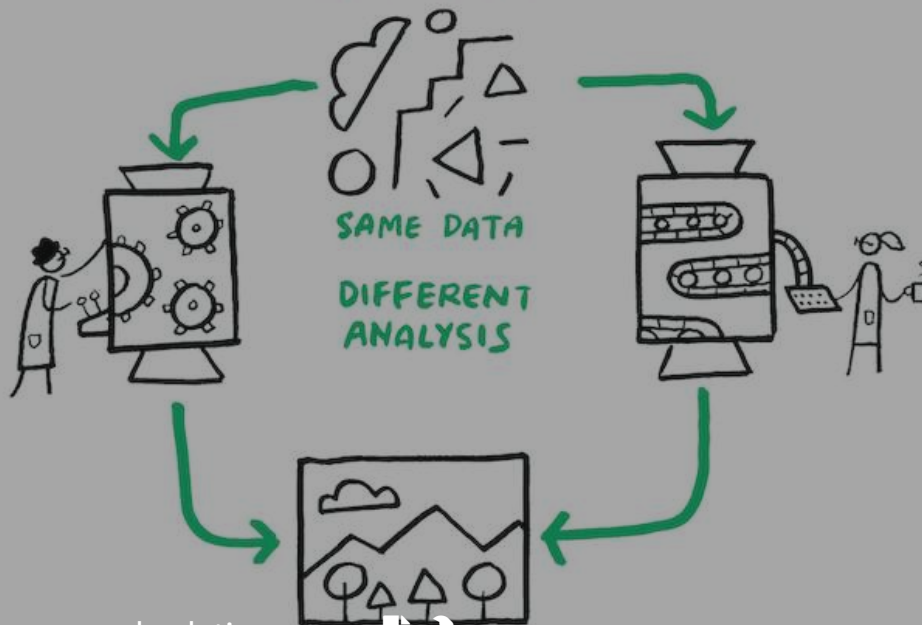
Scriberia

Software Engineering

for research data analysis

- **Reproducibility**: automation, syntax
- **Transparency**: docs, tests, rev control
- **Modularity**: packages and functions
- **Generalisability**: compartmentalise quirks
  - multiple-select questions
  - non-standard user-defined missing
  - erroneous coding of responses
  - reordering and reverse coding of response levels