UWE Bristol | University of the West of England

Building trust in research

Reproducibility, verification and transparency

Data Research Access and Governance Network

uwedragon.org

SDAP meeting on reproducibility

8th June 2021

Felix Ritchie

Access modes

- Open data
  - eg internet downloads

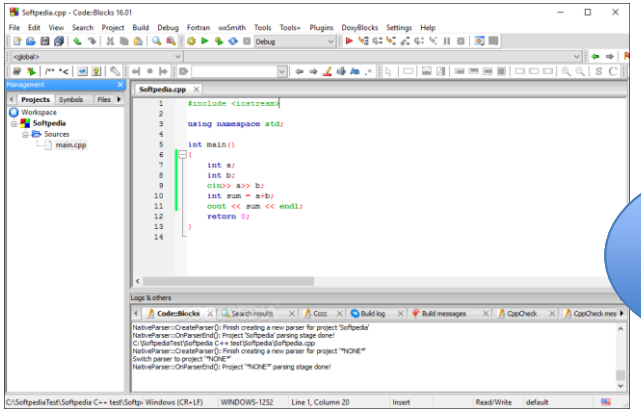
- Open access
  - Distributed data: eg UK Data Archive EUL
  - Ideal for easily de-identifiable social data
  - RDCs eg Secure Research Service
  - Ideal for business, health, longitudinal data
  - RJSs eg microdata.no, OpenSafely
  - Alt. to RDCs where less need to view data

- Restricted access
  - private data stores
  - self-collected data (where funders don't enforce access requirements)

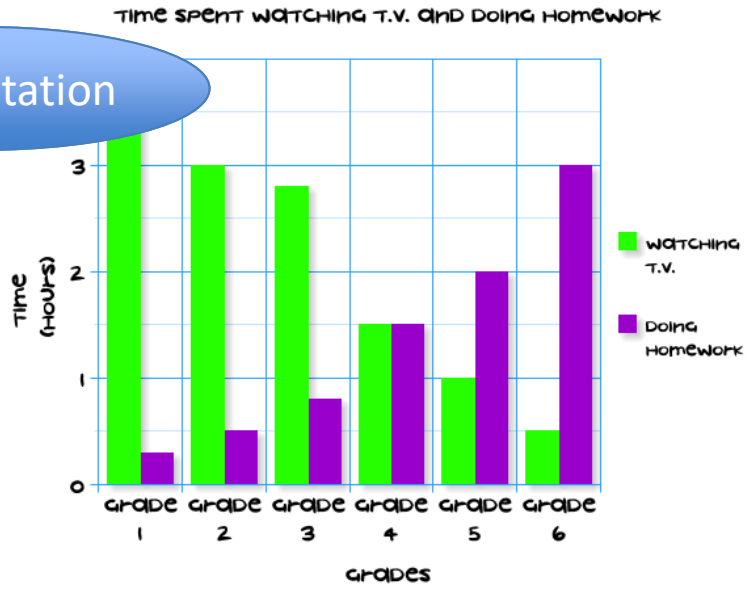
More valuable
Less traceable

Where does research come from?

Data



Code/ analysis

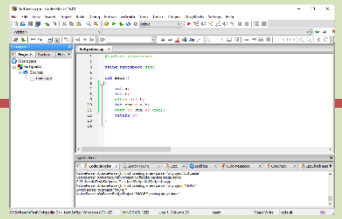


Interpretation

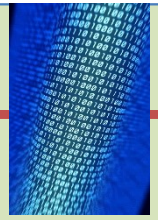
Presentation

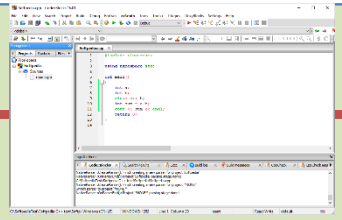
What makes reliable evidence?

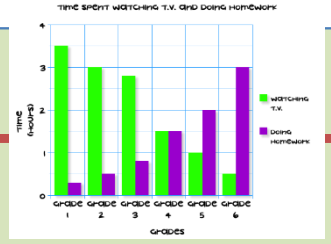
data verification



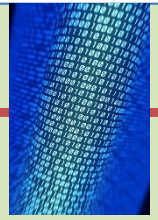
data and code verification

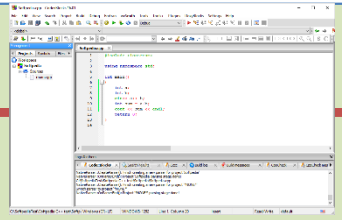


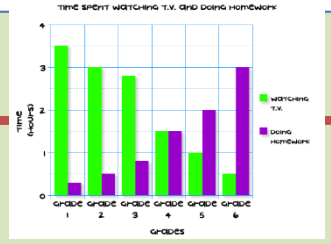




Research reproducibility







Research validation

Data Code Interpretation Presentation

Data verification

- What version of the data is used?
  - Vintage (esp for repeated surveys/cohort studies)
  - Subsets/linked data
  - Revisions
  - …
- Problem: how do users know version number?
- But doable eg DOIs, INEXDA annodata, RDA report
  - https://datascience.codata.org/articles/10.5334/dsj-2021-012/
- What about researcher-created data? ➔

Code verification

- Ideally, complete coding from source data to output
  - Including intermediate datasets
  - Advantage to data holders: no archiving of user data
- Slightly tricky…
  - requires discipline on the part of the researcher
  - Often not a linear path from input to output
  - Multiple researchers working on projects
- But doable and enforceable in TREs

Research reproducibility

- Interpretation (eg Excel) not coded as such
  - Needs version control
- Manual intervention to get coded outputs into interpretation engines
- Much more tricky…
  - Excel etc treated as 'live' interpretation engine
  - Often lots of extraneous information generated
- Doable, but with lots of scope for error

Research validation

- Central function of the peer-review process
  - But: mostly limited to what methodology section says
    - eg Reinhart and Rogoff
- Should it include data/code validation?
  - American Economic Review trying to validate all analysis
  - CASC offering this as a service
- What if you disagree with the method?
  - When does validation becomes approval?
- Limited objectivity – limited value?

Incentives to validate

- Researcher: self-validation
  - workflow management
  - credibility of research
  - journals
  - ethics
- Access managers: validation as a service
  - eg MedConf NHS TRE reports
  - credibility of environment
  - value of environment
  - internal audit
  - data management/archiving
  - commercial value

Disincentives to validate

- Researcher:
  - Time/effort
  - Discipline
- Access managers
  - Time/effort
  - Skills
  - Lack of co-operation from researchers
- Reviewers
  - Time/effort
  - Legal restrictions

Optimal model?

- Complete data versioning by data holder/manager
- + data & code validation by (researcher?)
- + publication of raw (coded) outputs?
- + separate peer review?

- Trust+transparency-based model?