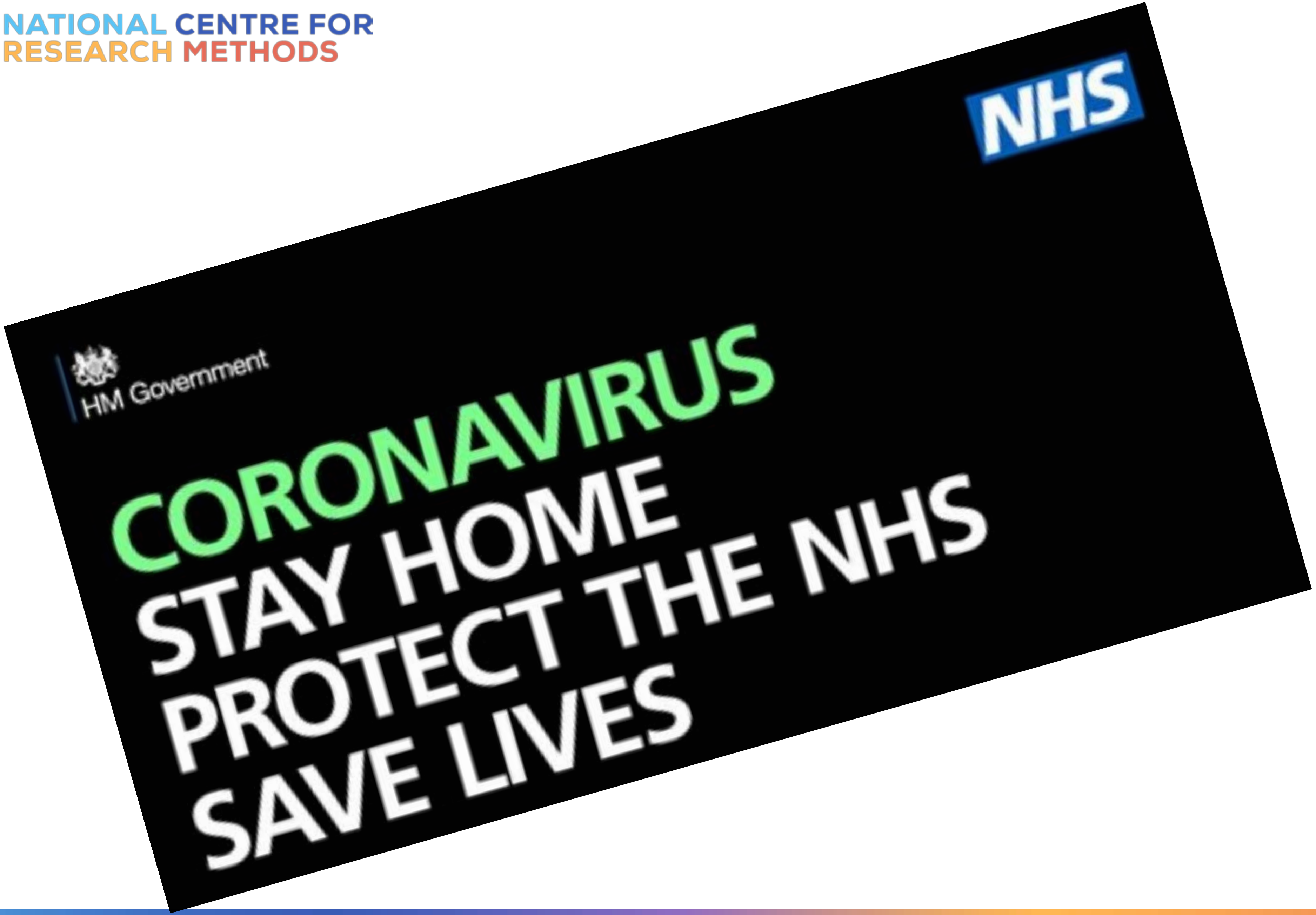
NATIONAL CENTRE FOR RESEARCH METHODS

The Data Analysis Workflow

Professor Vernon Gayle
vernon.gayle@ed.ac.uk
@Profbigvern
https://github.com/vernongayle

Economic and Social Research Council
UNIVERSITY OF Southampton
MANCHESTER 1824 The University of Manchester
THE UNIVERSITY OF EDINBURGH



HM Government

NHS

CORONAVIRUS
STAY HOME
PROTECT THE NHS
SAVE LIVES

The Data Analysis Workflow

A Thought Experiment (be honest…)

1. Have you ever lost a file?

2. Have you ever wondered if you have deleted a file?

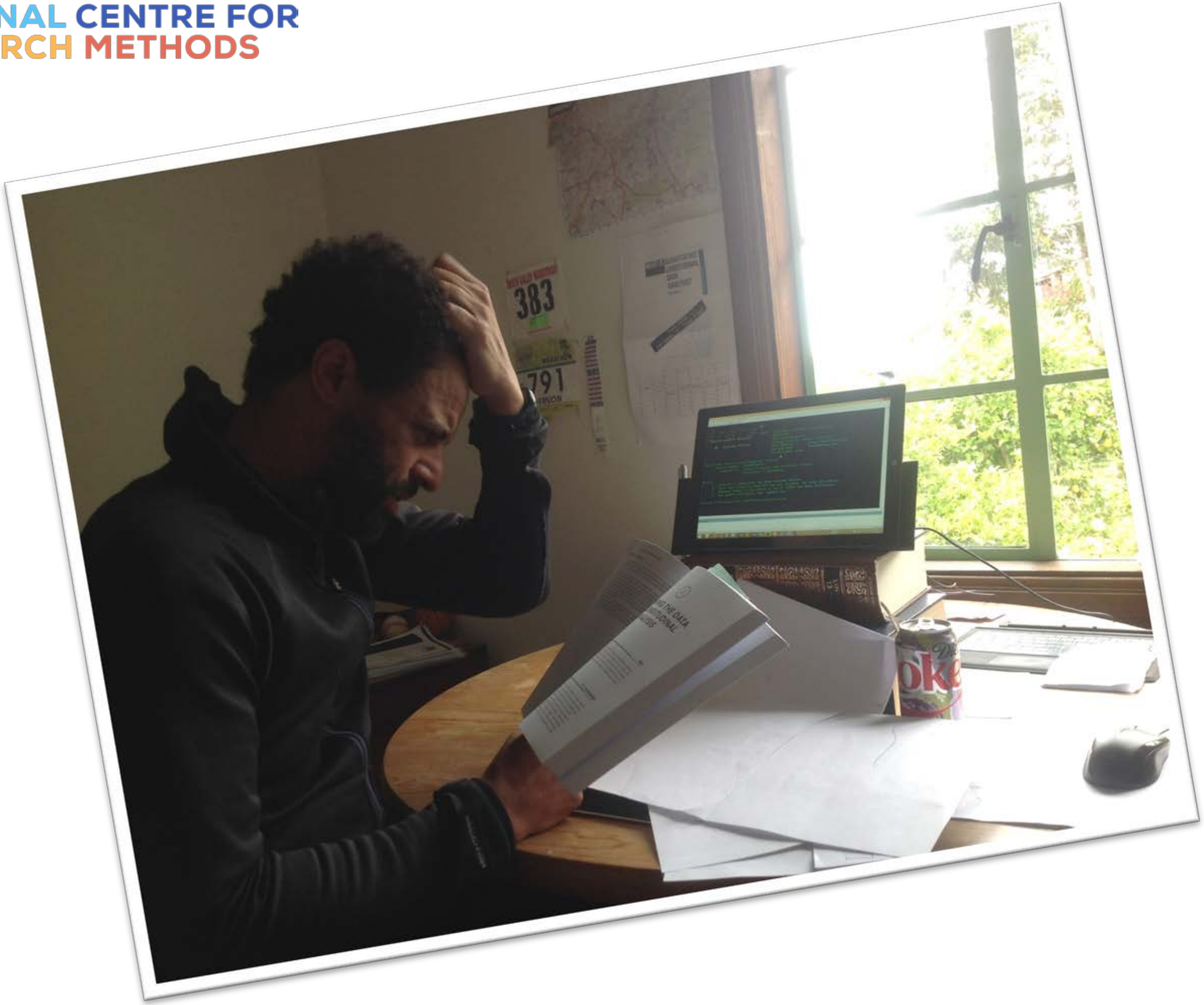
3. Have you and a colleague ever been working on different versions of a file?

A Thought Experiment (be honest…)

4. Have you ever struggled to identify data files?

chapter1_2019.dat

chap1_2019.dat


383
791

Volume 85 • Number 1 • February 2020

AMERICAN SOCIOLOGICAL REVIEW

2019 ASA Presidential Address

Sociology Engaged in Social Justice
Mary Romero

Organizational Practices and the Reproduction of Racial Inequality

Getting In, Getting Hired, Getting Sideways Looks: Organizational Hierarchy and Perceptions of Racial Discrimination
Adia Harvey Wingfield and Koji Chavez

Mano Suave–Mano Dura: Legitimacy Policing and Latino Stop-and-Frisk
Victor M. Rios, Greg Prieto, and Jonathan M. Ibarra

The Role of Ambiguity and Ambivalence in Social Life

Explaining Ambiguity: A Moral Polysemy Approach to Variation in Economic Practices
Guillermina Altomonte

Vaccine Refusal and Pharmaceutical Acquiescence: Parental Control and Ambivalence in Managing Children's Health
Jennifer A. Reich

The Value of Education and Skills in Comparative Perspective

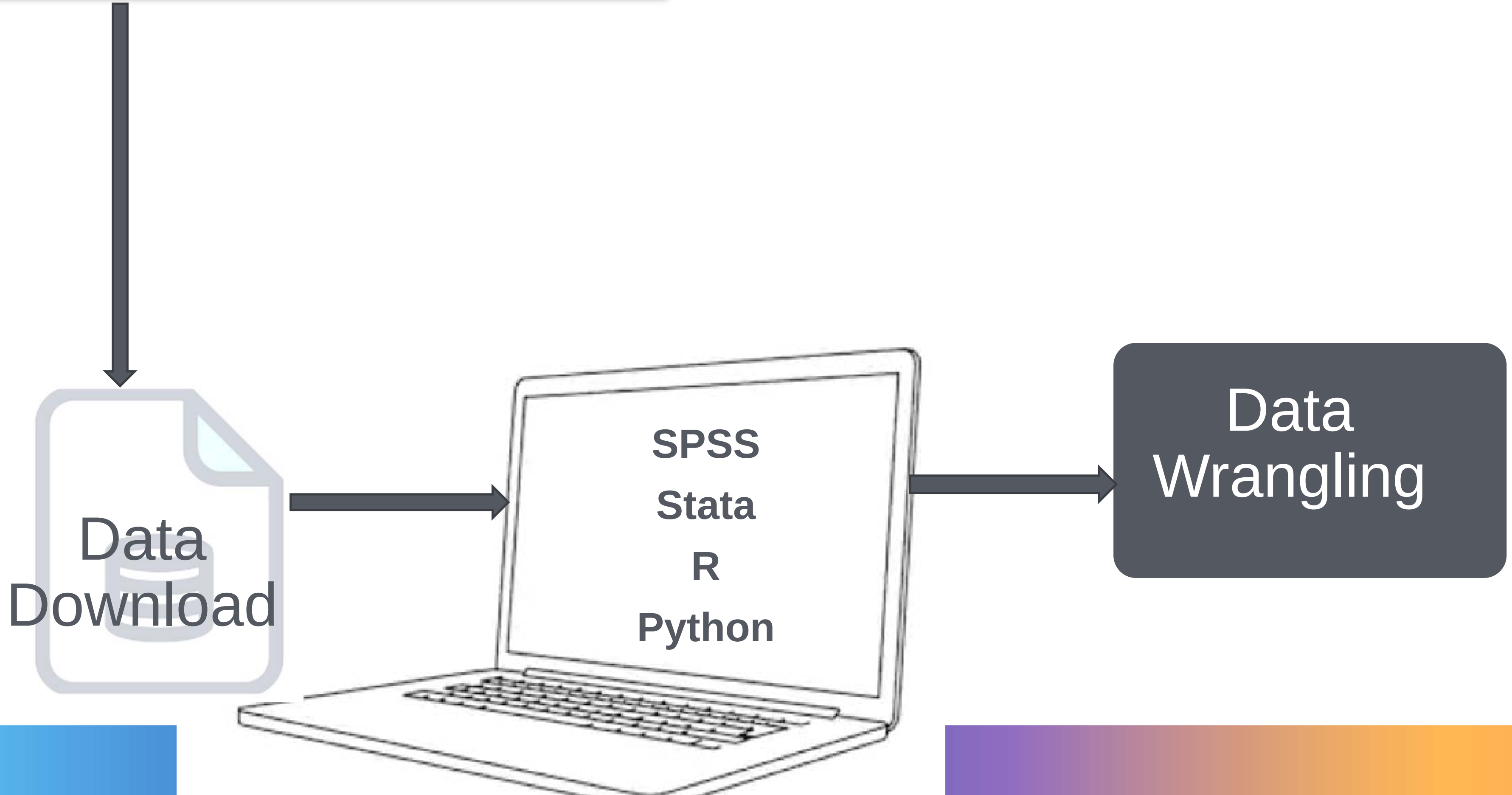
Educational Expansion, Skills Diffusion, and the Economic Value of Credentials and Skills
Satoshi Araki

Comment and Reply

Comment on Desmond, Papachristos, and Kirk's "Police Violence and Citizen Crime Reporting in the Black Community"
Michael Zoorob

Reply to Zoorob
Matthew Desmond, Andrew V. Papachristos, and David S. Kirk

• A Journal of the American Sociological Association •



Data Download

SPSS

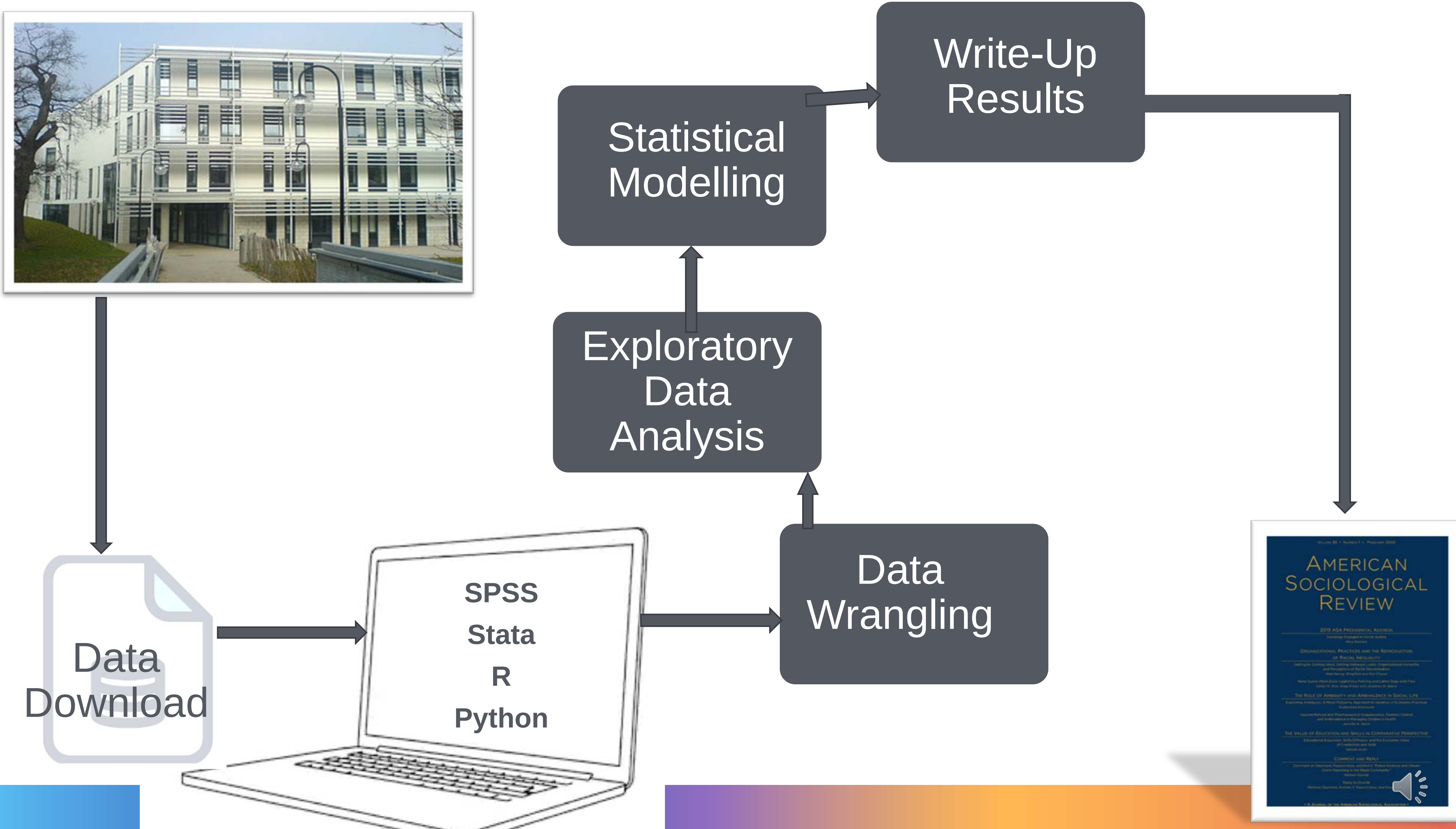
Stata

R

Python

Data Wrangling

AMERICAN SOCIOLOGICAL REVIEW



Data Download

SPSS

Stata

R

Python

Data Wrangling

Exploratory Data Analysis

Statistical Modelling

Write-Up Results

The Workflow

Analysing data without a planned and organised workflow can be compared to drinking and driving. In both situations it doesn’t matter how careful you are, it is still highly likely to end in a wreck!

(Philip Stark UC Berkeley)

Therefore just like drinking and driving, we strongly warn against not having a systematic workflow

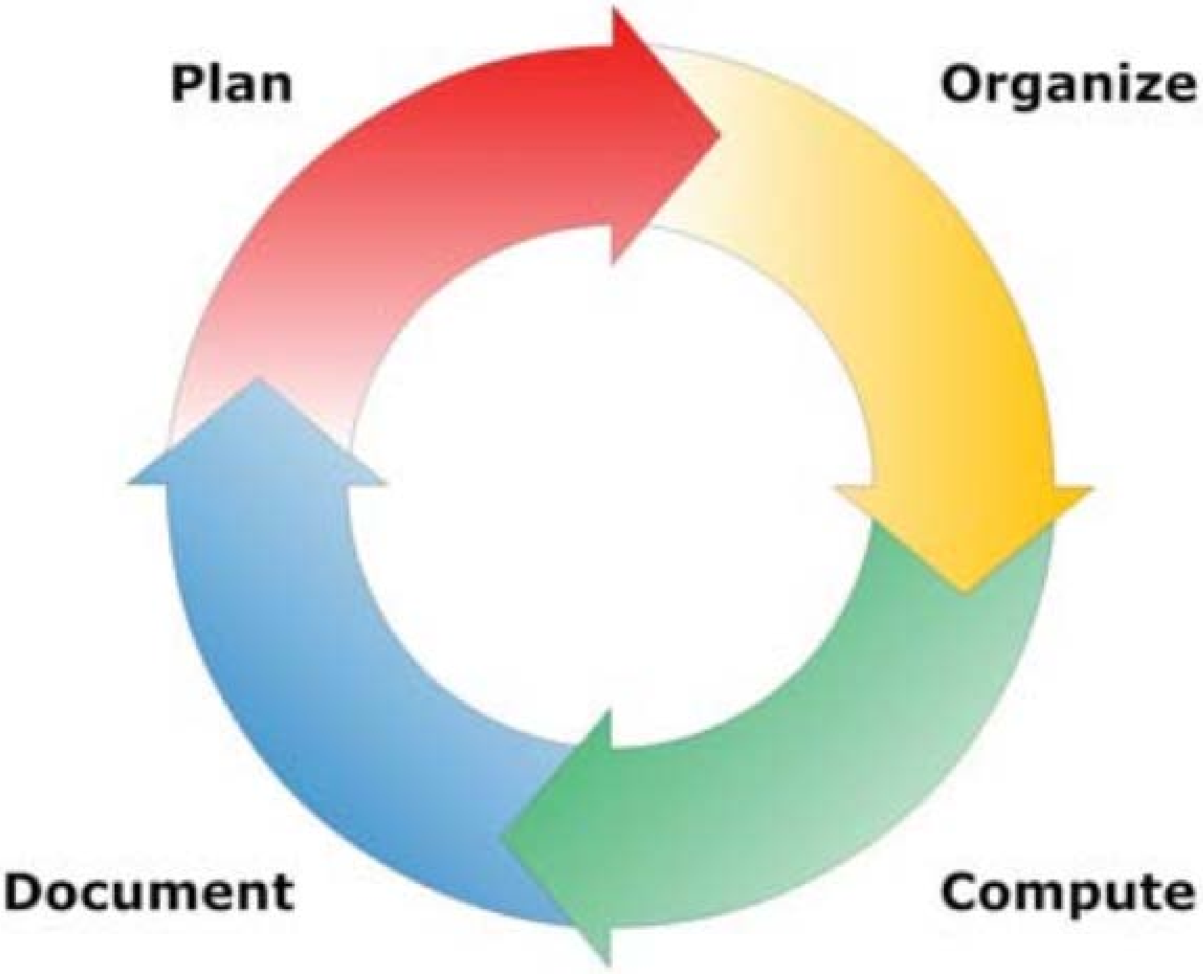
The Workflow

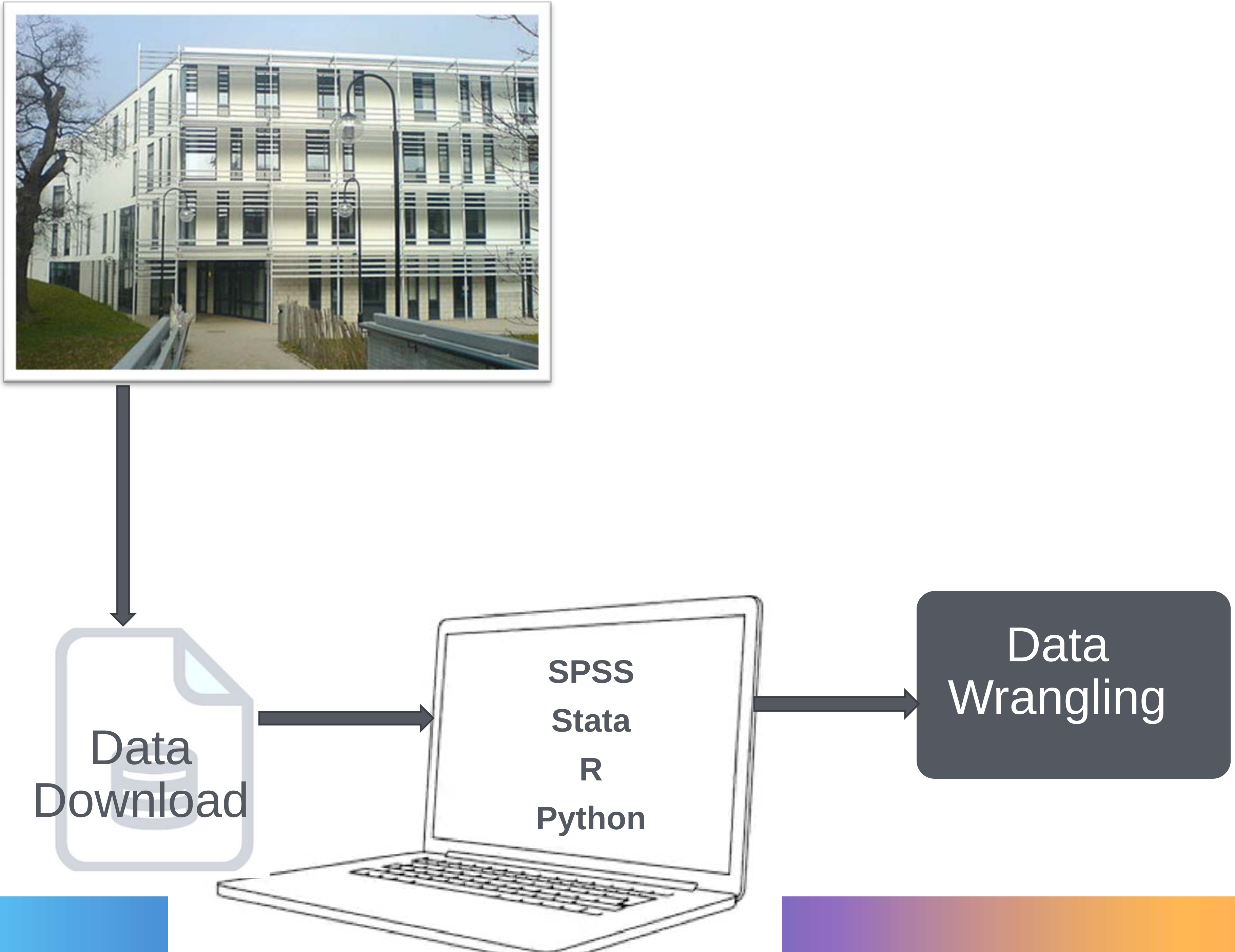
Workflow should be planned and carefully orchestrated

Workflow MUST NOT be *adhoc*

(e.g. piece-meal, developed as a reaction to mistakes etc.)

The Workflow Cycle





Data Download

SPSS

Stata

R

Python

Data Wrangling

Drop down menus = no audit trail

GUIs will leave you in a sticky mess!

Data Wrangling

- Selecting variables (surveys have large *k*)
- Operationalising measures (e.g. income; social class; education)
- Re-coding variables
- Selection cases (sub-samples in large surveys)
- Missing data

Data Wrangling

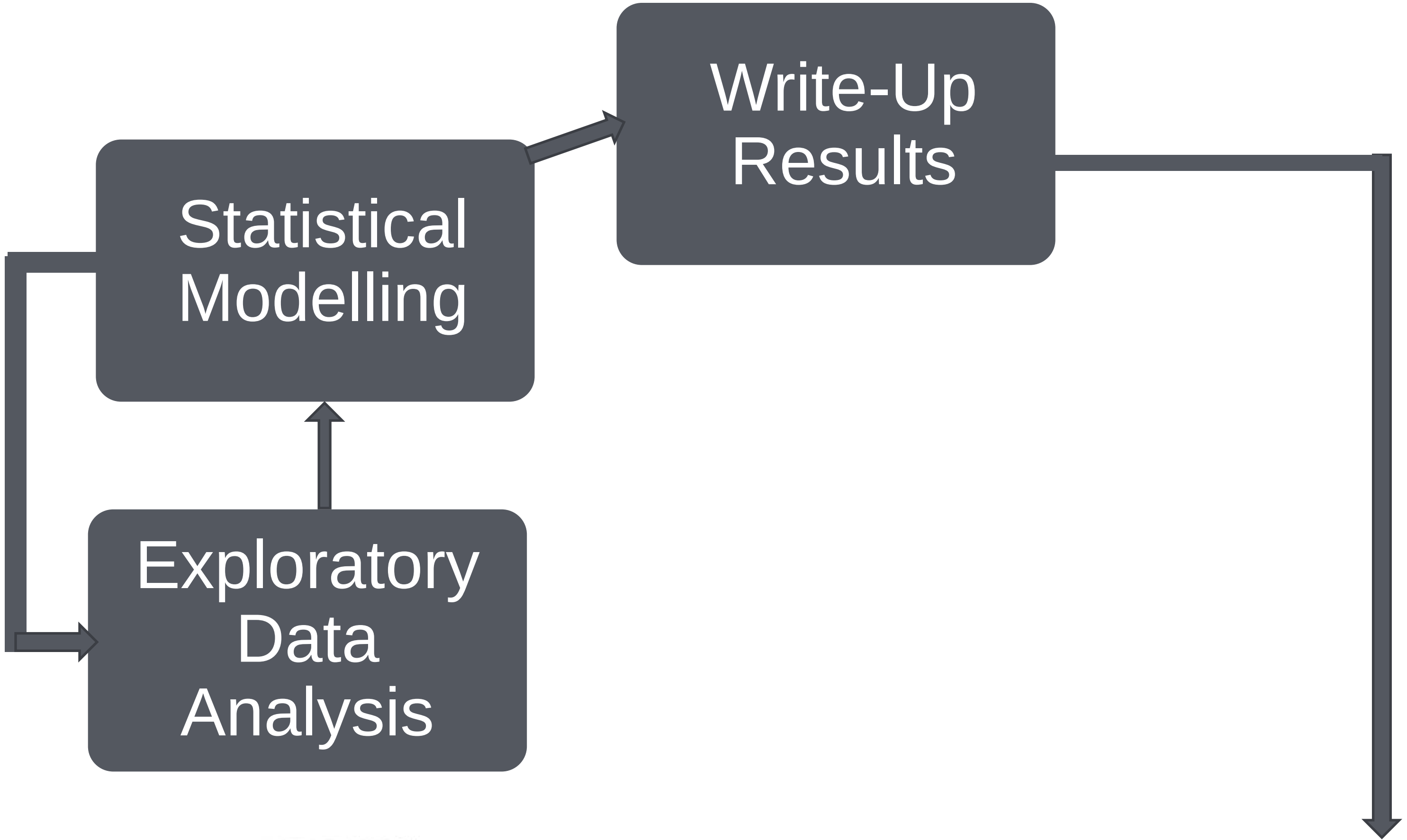
- Selecting variables (surveys have large *k*)
- Operationalising measures (e.g. income; social class; education)
- Re-coding variables
- Selection cases (subsamples in large surveys)
- Missing data

CAN THESE ACTIONS BE TRACED IN MY AUDIT TRAIL

Minor actions in the

Data Wrangling Phase

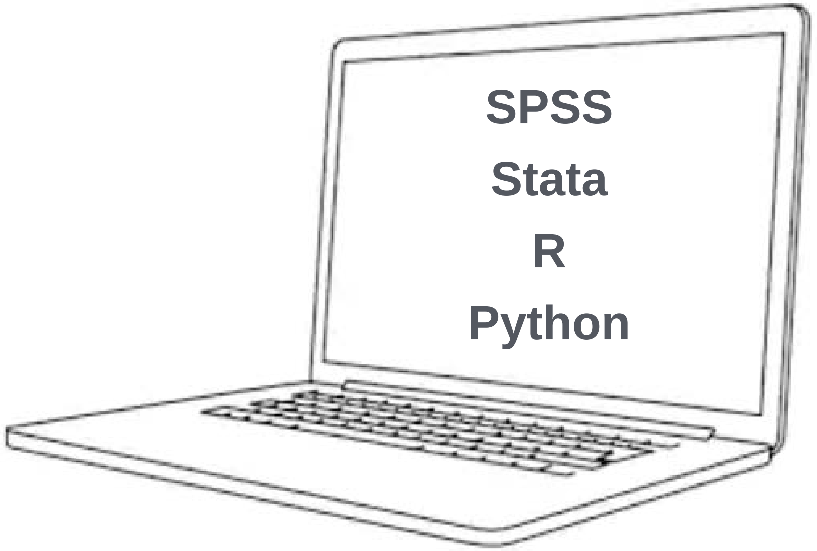
can have major consequences!



Exploratory Data Analysis

Statistical Modelling

Write-Up Results



SPSS

Stata

R

Python

AMERICAN SOCIOLOGICAL REVIEW

2019 ASA PRESIDENTIAL ADDRESS

ORGANIZATIONAL PRACTICES AND THE REPRODUCTION OF RACIAL INEQUALITY

THE ROLE OF AMBIGUITY AND AMBIVALENCE IN SOCIAL LIFE

THE VALUE OF EDUCATION AND SKILLS IN COMPARATIVE PERSPECTIVE

COMMENT AND REPLY

Data Analysis

- Which cases
- Which variables (format)
- Missing data
- Estimation method (mle; gls)
- Weights and survey structure (svy)
- Setting a seed
- Number of quadrature points
- Which version or library

Data Analysis

- Which cases
- Which variables (format)
- Missing data
- Estimation method (mle; gls)
- Weights and survey structure (svy)
- Setting a seed
- Number of quadrature points
- Which version or library

CAN THESE ACTIONS BE TRACED IN MY AUDIT TRAIL

Improving the Workflow

Better supporting YOU and what YOU DO

Not changing you into something YOU ARE NOT

ALL SERIOUS WORK MUST

BE REPRODUCIBLE!

There MUST be an audit trail

A Planned Workflow Has Benefits

Accuracy

Progamming Efficiency

Transparency

Reproducibility

Four Pillars of Wisdom

- Accuracy
  - minimising information loss and errors in analyses and output
- Programming Efficiency
  - automation, maximising features in software
- Transparency
  - showing what you did, why, when, how
- Reproducibility
  - same results every time whoever or wherever
  - editing, rewriting reports or re-submission of papers

Long’s Law

It is always easier to document today than it is tomorrow!

Corollary 1:
Nobody likes to write documentation

Corollary 2:
Nobody ever regrets having written documentation

Long’s Law

Has anyone in the history of data analysis ever said

“these files are too well documented”

Good News

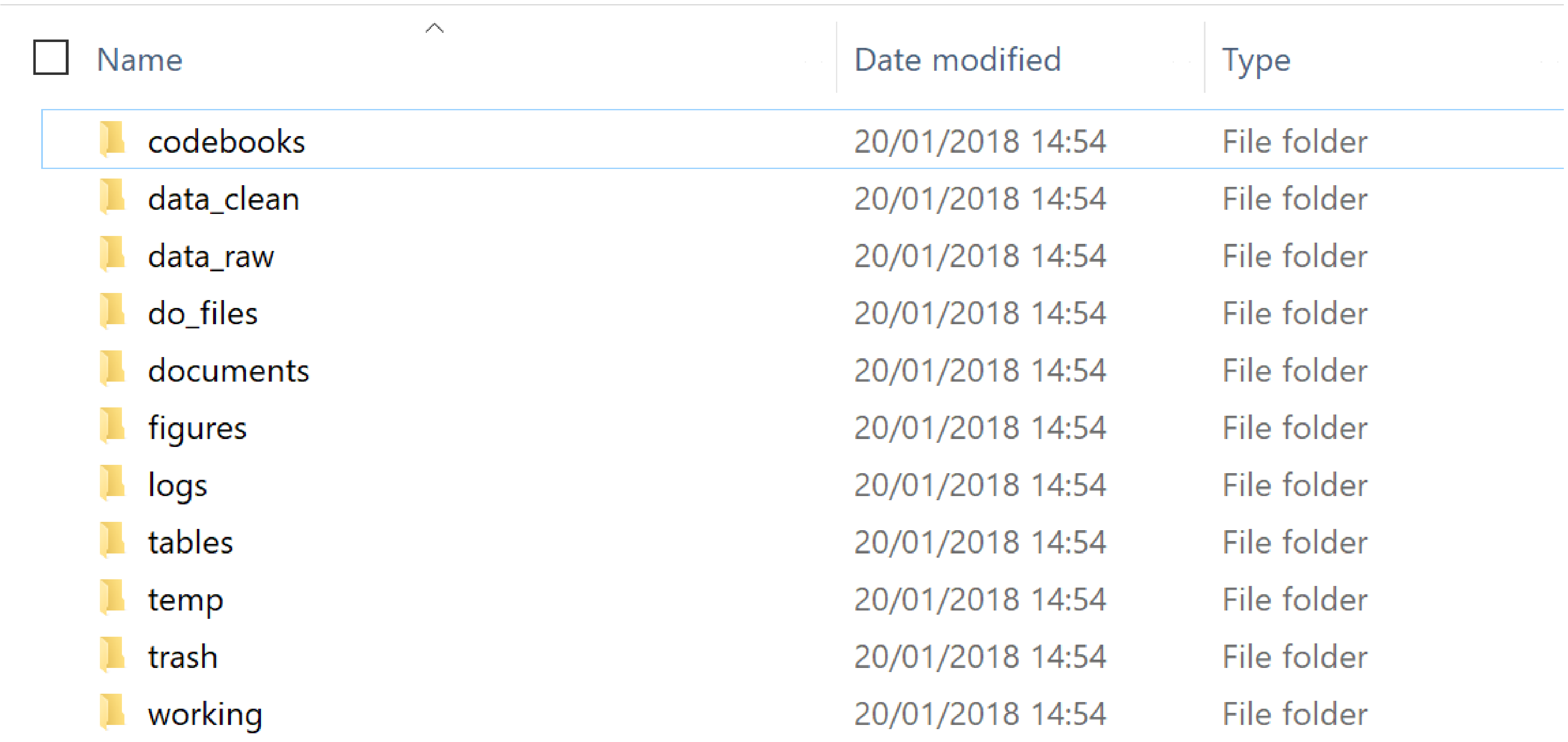
- Improving the workflow with a modest amount of effort

- The less experience you have the better

  (start from the very beginning)

Some Practical Strategies and Tactics

Standard Directory Structure



□ Name	Date modified	Type
codebooks	20/01/2018 14:54	File folder
data_clean	20/01/2018 14:54	File folder
data_raw	20/01/2018 14:54	File folder
do_files	20/01/2018 14:54	File folder
documents	20/01/2018 14:54	File folder
figures	20/01/2018 14:54	File folder
logs	20/01/2018 14:54	File folder
tables	20/01/2018 14:54	File folder
temp	20/01/2018 14:54	File folder
trash	20/01/2018 14:54	File folder
working	20/01/2018 14:54	File folder

File Naming Protocols

File Name = name_date_depositor's initials_version_type

Therefore **bhpsaindresp_20140506_vg_v1.dta**

Would be a

a. The British Household Panel Survey File “aindresp”

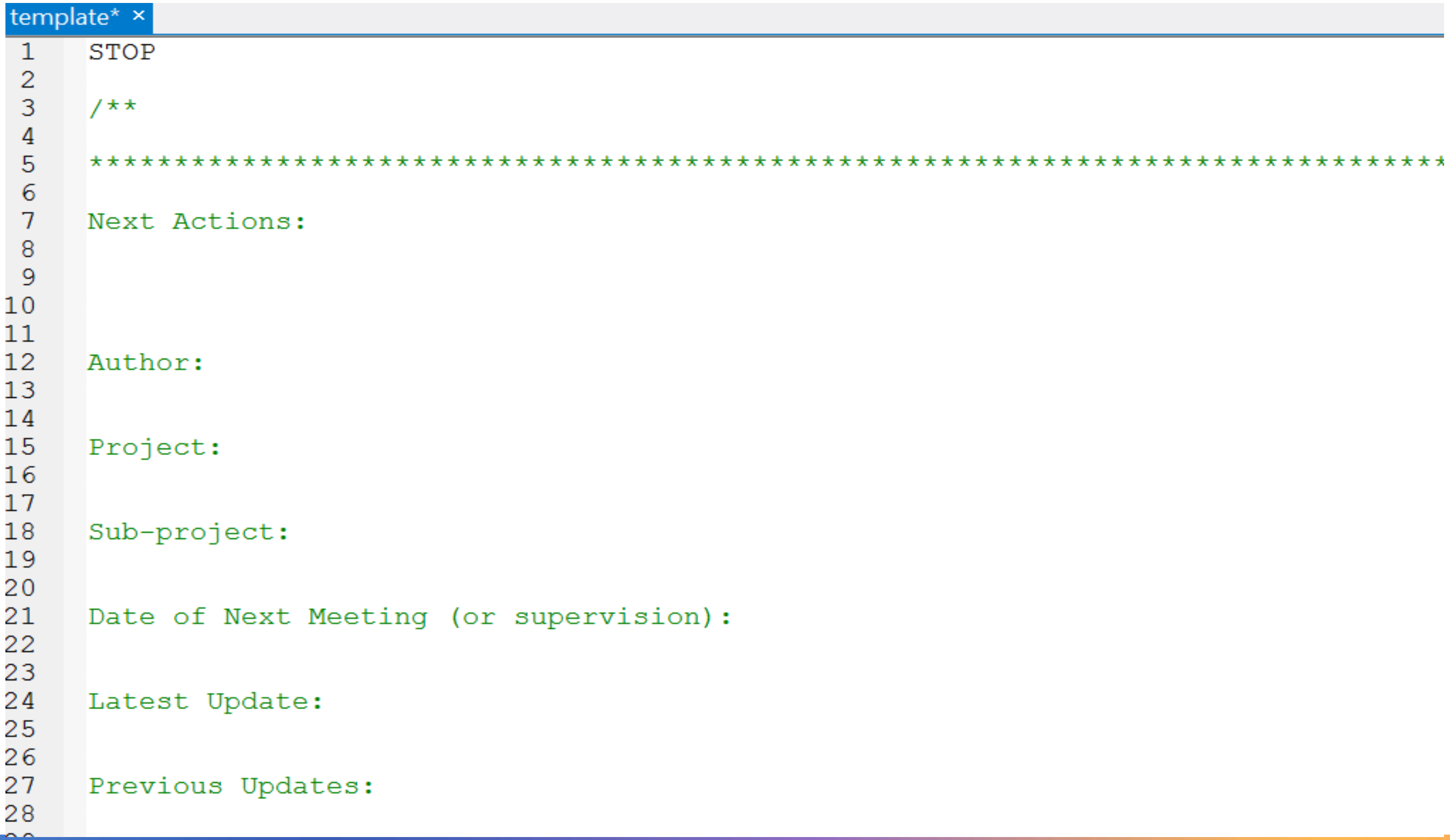
b. Deposited on 6th May 2014

c. Deposited by vg (Vernon Gayle)

d. Version v1

e. File type (e.g. a Stata .dta file)

Analysis File Structure and Information



```
STOP

/**

******************************************************************************************

Next Actions:




Author:


Project:


Sub-project:


Date of Next Meeting (or supervision):


Latest Update:


Previous Updates:

```

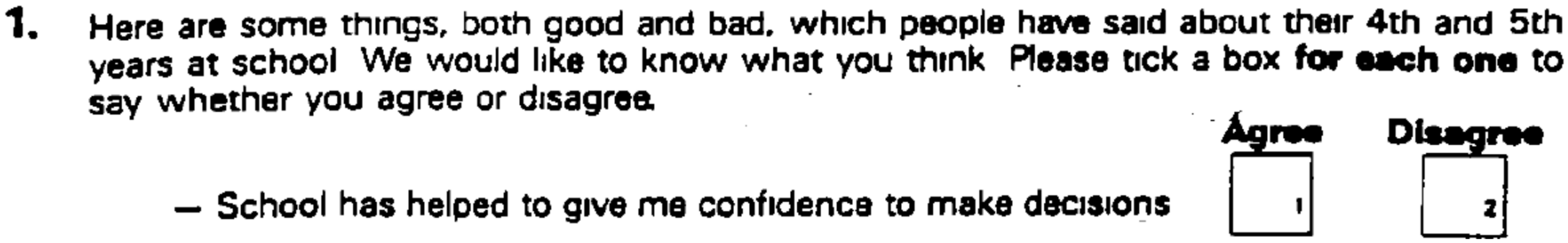
Variable Naming Protocols

- Good data providers will tend to stick to clearly defined variable naming conventions

- Wave 1 of UKHLS ***a_sex*** is the gender variable, and in wave 2 of the study ***b_sex*** is the gender variable (UK Data Archive Study Number 6614)

- Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 the variable ***t0schtyp*** is the type of school that the pupil attended in Year 11 (at time point t0 in the study) (UK Data Archive Study Number 5765)

Variable Naming Protocols



1. Here are some things, both good and bad, which people have said about their 4th and 5th years at school We would like to know what you think Please tick a box **for each one** to say whether you agree or disagree.

	Agree	Disagree
— School has helped to give me confidence to make decisions	□ 1	□ 2

By contrast in the less well curated Youth Cohort Study of England and Wales Cohort 4 (UK Data Archive Study Number 3107) this is the opaque variable ***dx11_a***

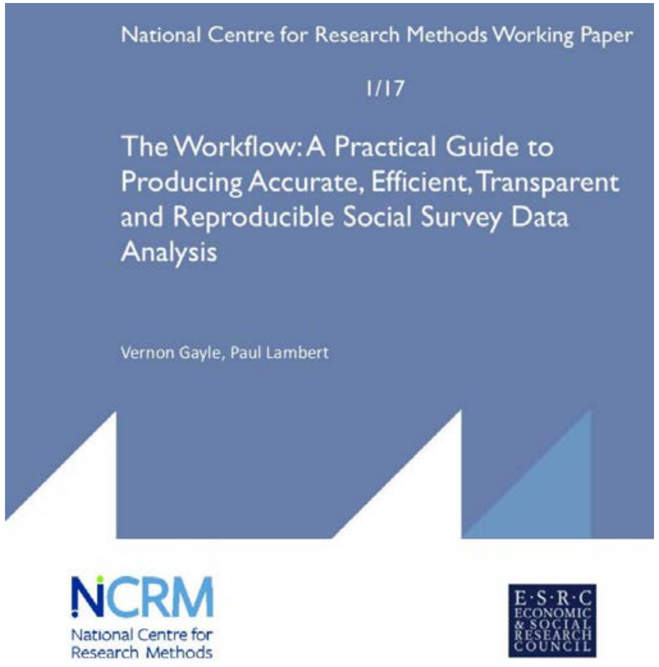
Estimating Work Time…

Benefits of a Good Workflow

- Limits duplication of effort
- Systematic work minimises errors
- Helps to detect errors
- Editing work, resubmitting papers etc.
- Aids collaboration
- Supports additional (e.g. future) work

What To Do…

- Audit trail
- Don’t use drop menus (a gui = a sticky mess)
- Systematic directory structure, file names, versions, variable names etc.
- Plan! don’t do *ad hoc* work or work on the fly
- Write comments


National Centre for Research Methods Working Paper

1/17

The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis

Vernon Gayle, Paul Lambert

NCRM National Centre for Research Methods

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

http://eprints.ncrm.ac.uk/4000/

J. Scott Long has posted a really good pdf version of a talk on the workflow http://www.ihrp.uic.edu/files/Workflow%20Slides%20JSLong%20110410.pdf

ALL SERIOUS WORK MUST

BE REPRODUCIBLE!

There MUST be an audit trail

How to cite this video

Gayle, V. (2020) *The Data Analysis Workflow.* Available at: https://www.ncrm.ac.uk (Accessed: day month year)

NATIONAL CENTRE FOR RESEARCH METHODS

The Data Analysis Workflow

Professor Vernon Gayle

vernon.gayle@ed.ac.uk

@Profbigvern

https://github.com/vernongayle

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH