

---

# What does a user look for in a survey dataset?

(What does this user look for in a survey dataset)

(A few things that this user would like you to think about)

---

Vernon Gayle, FRSE FAcSS  
(University of Edinburgh)

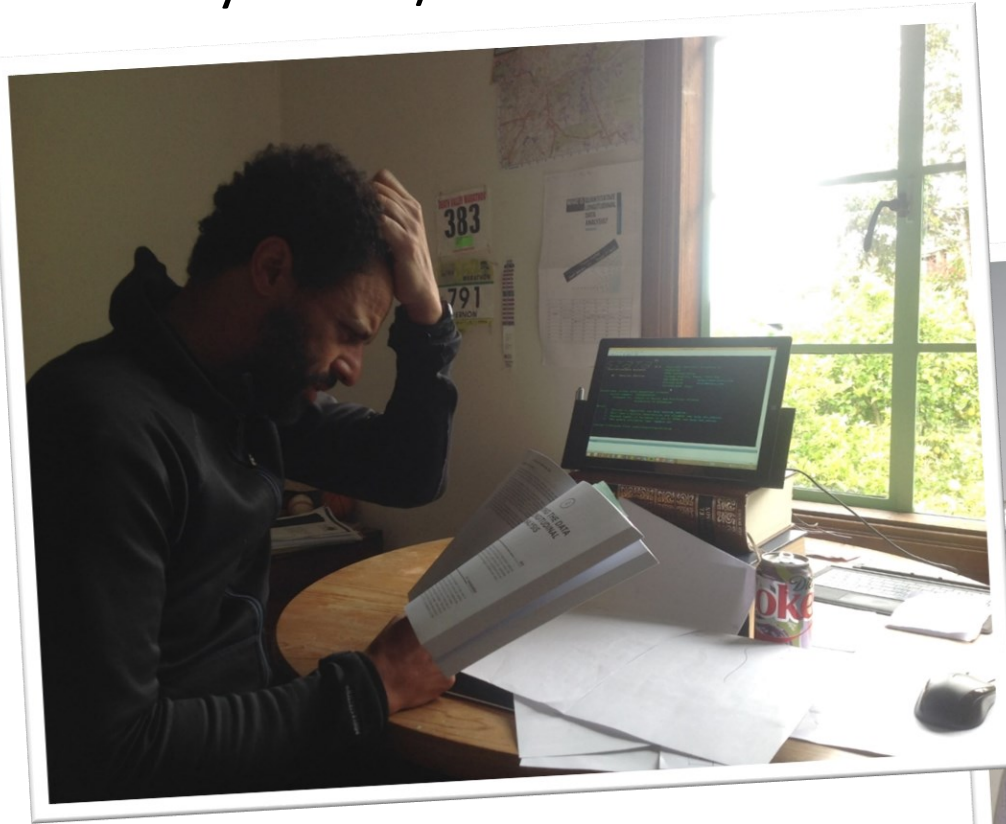
**Challenges and Opportunities for Social Survey Data Collection in Scotland**

**Survey Practice Forum**

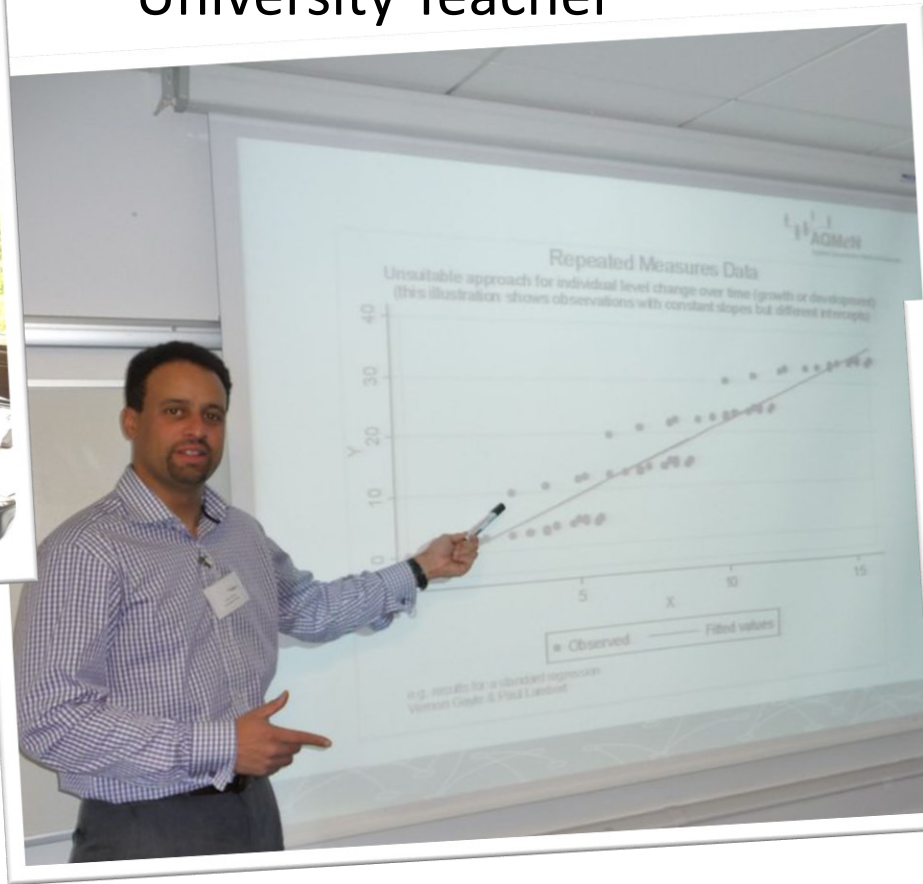
**Royal Society of Edinburgh**

6<sup>th</sup> December 2024

Survey Users / Researcher



University Teacher



Postgraduate Supervisor



# Things have improved for survey data analysts...



1994

new datasets,  
the internet,  
faster computers,  
better storage,  
improved software,  
better documentation,  
some linked admin data



2024



# What does a user look for in a survey dataset?





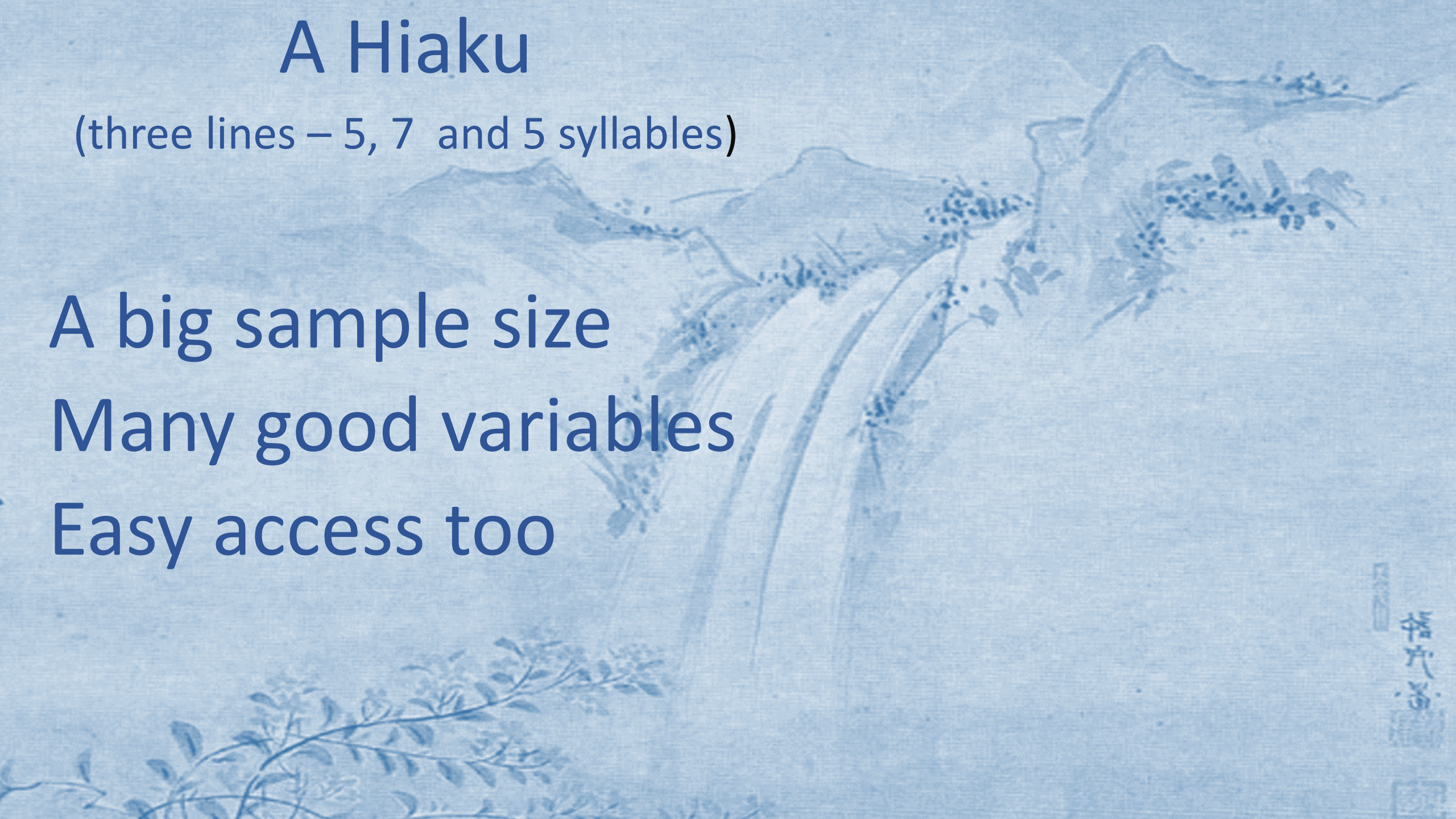
# A Hiaku

(three lines – 5, 7 and 5 syllables)

A big sample size

Many good variables

Easy access too





# A Hiaku

(three lines – 5, 7 and 5 syllables)

A big sample size

Many good variables

Easy access too

Big  $n$

Large  $k$

Multiple  $t$

Suitable  $j$  (geography)

# RRS SIR DAVID ATTENBOROUGH

Once you have set eyes on the RRS Sir David Attenborough, you won't forget her. Measuring in at 129 metres, the ship is as long as 10 buses and weighs 10,400 tonnes - that's 1,400 elephants. Built by Cammell Laird to a Rolls Royce design and kitted out with state-of-the-art facilities, the ship will push the boundaries of polar science and exploration.

It is made up of 1 MILLION pieces of steel, and contains over 30 KM of pipes and more than 750 KM of electric and data cables.

The ship has beds for 30 CREW and 60 SCIENTISTS and SUPPORT STAFF.

CTD (conductivity, temperature, and depth) - a collection of sensors deployed overboard to detect how the salinity (salt levels) and temperature of the water column change relative to depth.

Scientific winch system deploys equipment, such as rock drills, overboard.

Science crane

Main crane (DS tower)

Side & frame engines serve equipment overboard. Winch control room.

Crane

Satellite communications

Bridge

Crane

Helideck

Officer and crew cabins

Bar, lounge and mess room

Cargo tender "Barror" delivers people and supplies to land.

Hull designed to break through ice one metre thick.

LABORATORIES & WORKSPACES

There will be 14 laboratories on board and at least 10 sleeping containers with scientific equipment that can be reconfigured to keep up with changing technologies and techniques.

Work boat "Erebus" transports personnel and supplies.

ROCK DRILLS

Deployed from the stern, sides or moonpool of the ship, drills will sample soft sediment and rock up to 2000 metres underwater.

Shipping containers with scientific equipment

A 5m propeller

MOON POOL

Scientists can lower and raise equipment (such as ROVs) through the moon pool, a vertical hole running through the hull of the vessel. This makes it easier and safer to deploy scientific equipment in the rough polar oceans and ice-covered waters.

Electric propulsion motors

MOON POOL

ROV (remotely operated underwater vehicle)

MAINE ROBOTICS

The ship will act as a central platform for deploying state-of-the-art autonomous and remotely operated vehicles. These will explore untouched parts of the ocean and atmosphere. Remotely controlled vehicles will be connected to the ship and powered via a cable - just like an umbilical cord. Autonomous underwater vehicles, like the *Roamer* Midwater Autonomous Long Range, will have no link to the ship and will travel deep beneath ice shelves and at the edge of active glaciers.

LIVING ON BOARD

Scientists and crew will be able to unwind using the gym, sauna, bar, and TV facilities. They will sleep in a mixture of single and double-occupancy cabins.

Cabins are located away from the ship's bow to reduce the effects of motion.

ENGINE

The engines will run as silently as possible to avoid interference with the rigors of the ship's acoustic instruments, which use echo sounders to measure life in the water and map the sea floor.

*'Roamer' Midwater Autonomous Long Range*

UK Research and Innovation

NERC SCIENCE OF THE ENVIRONMENT

British Antarctic Survey

CAMMELL LAIRD

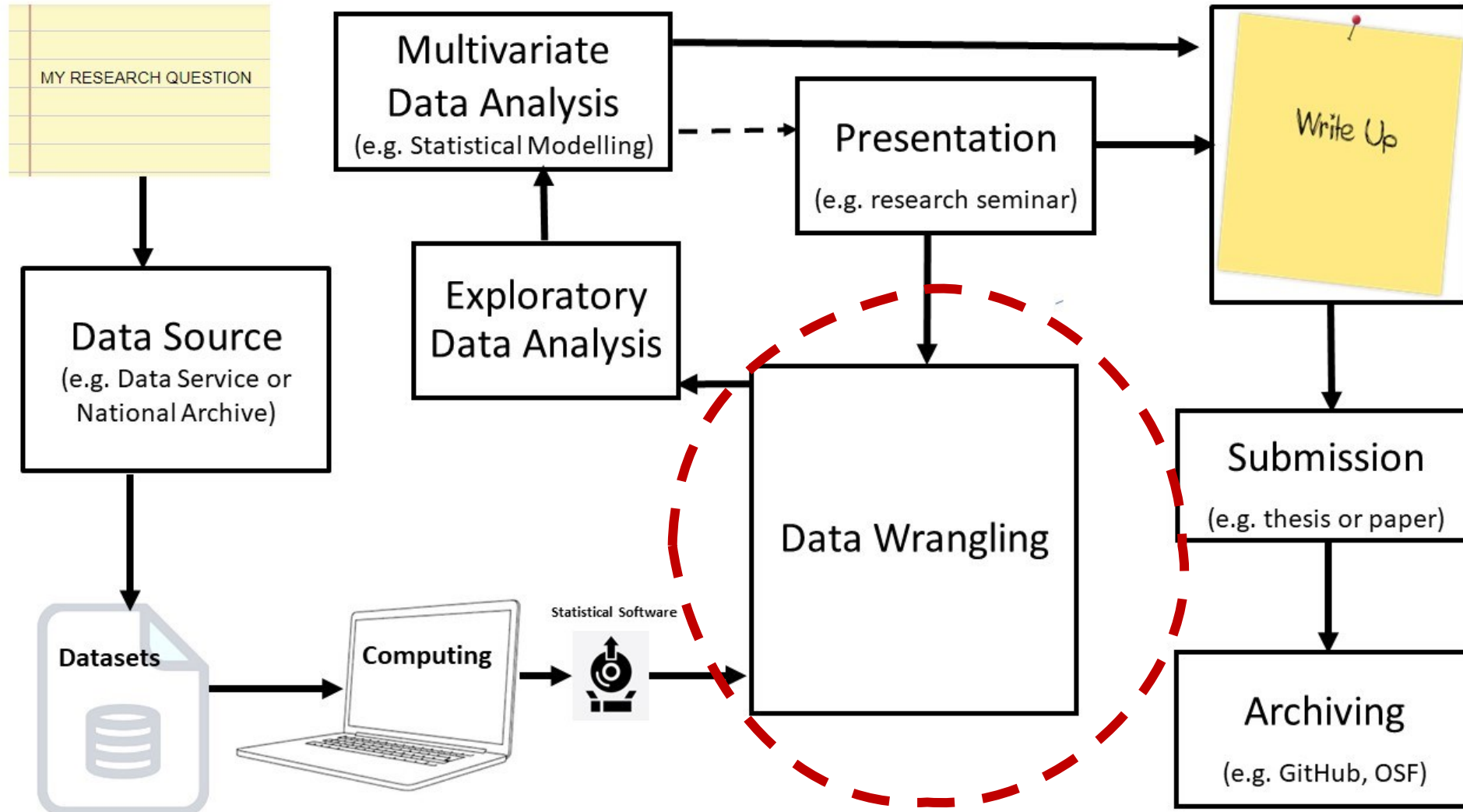
Copyright: UKRI, NERC, Ben Gilliland

## *Birth Cohorts*

## Annual Population Surveys

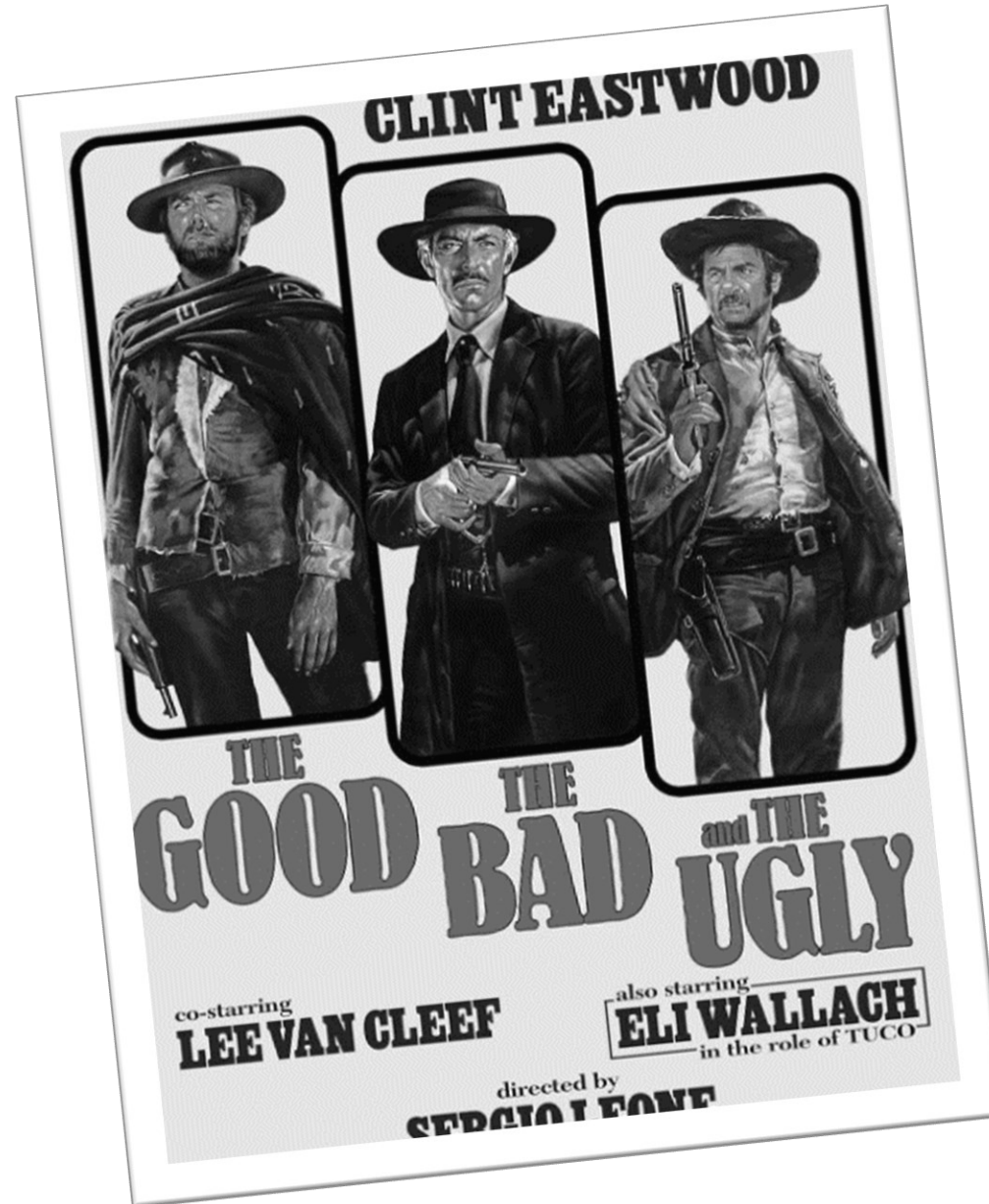
*Challenge to encourage users to be more tolerant –  
these surveys are not designed to address a specific hypothesis or for researching  
single topics or research field*

# Producing (nearly) Research Ready Datasets





# Complex Designs and Selection Strategies



## **The Good**

Assists fieldwork challenges

Provide territorial data (e.g. Scotland)

Facilitate the geographical analyses (e.g. areas)

Facilitate analyses of under-represented groups (e.g. ethnic minority)

## **The Bad**

Analysts ignoring survey designs and selection strategies

## **The Ugly**

Not adequately representing design in analysis

Not being able to provide some measures (e.g. model fitting statistics)

Missing data methods often trickier (e.g. svy m.i. models)



# Complex Designs and Selection Strategies

## **Challenges**

Are complex designs and selection strategies necessary?

Can some designs be simplified to help data analysts?

## **Opportunities**

Provide more training

Show routine examples

Show tricky examples (e.g. svy m.i. models)

# Documentation

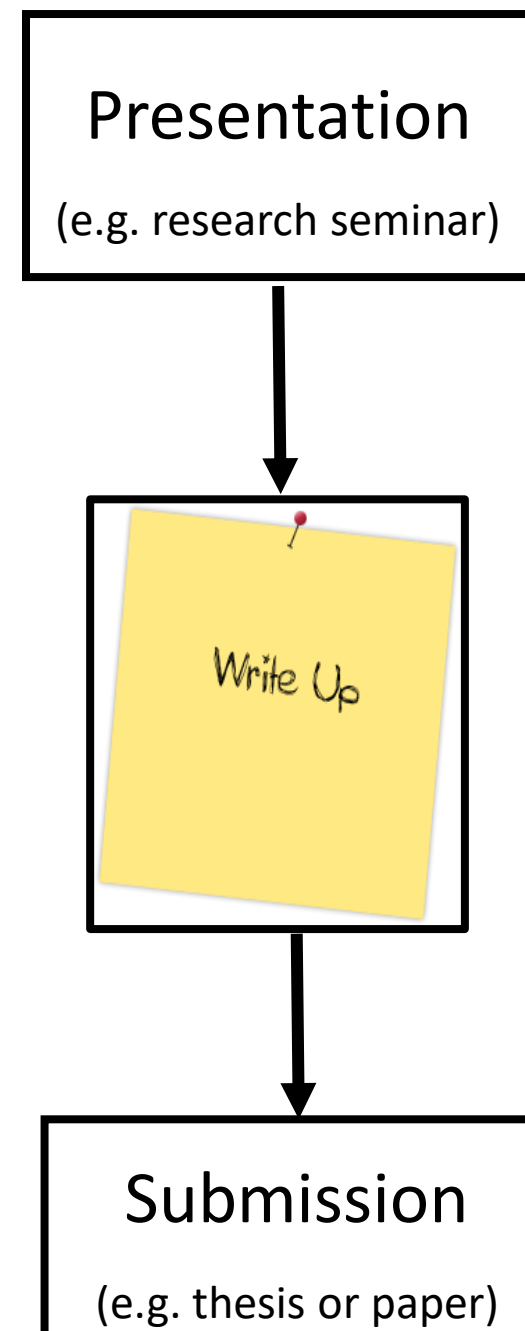
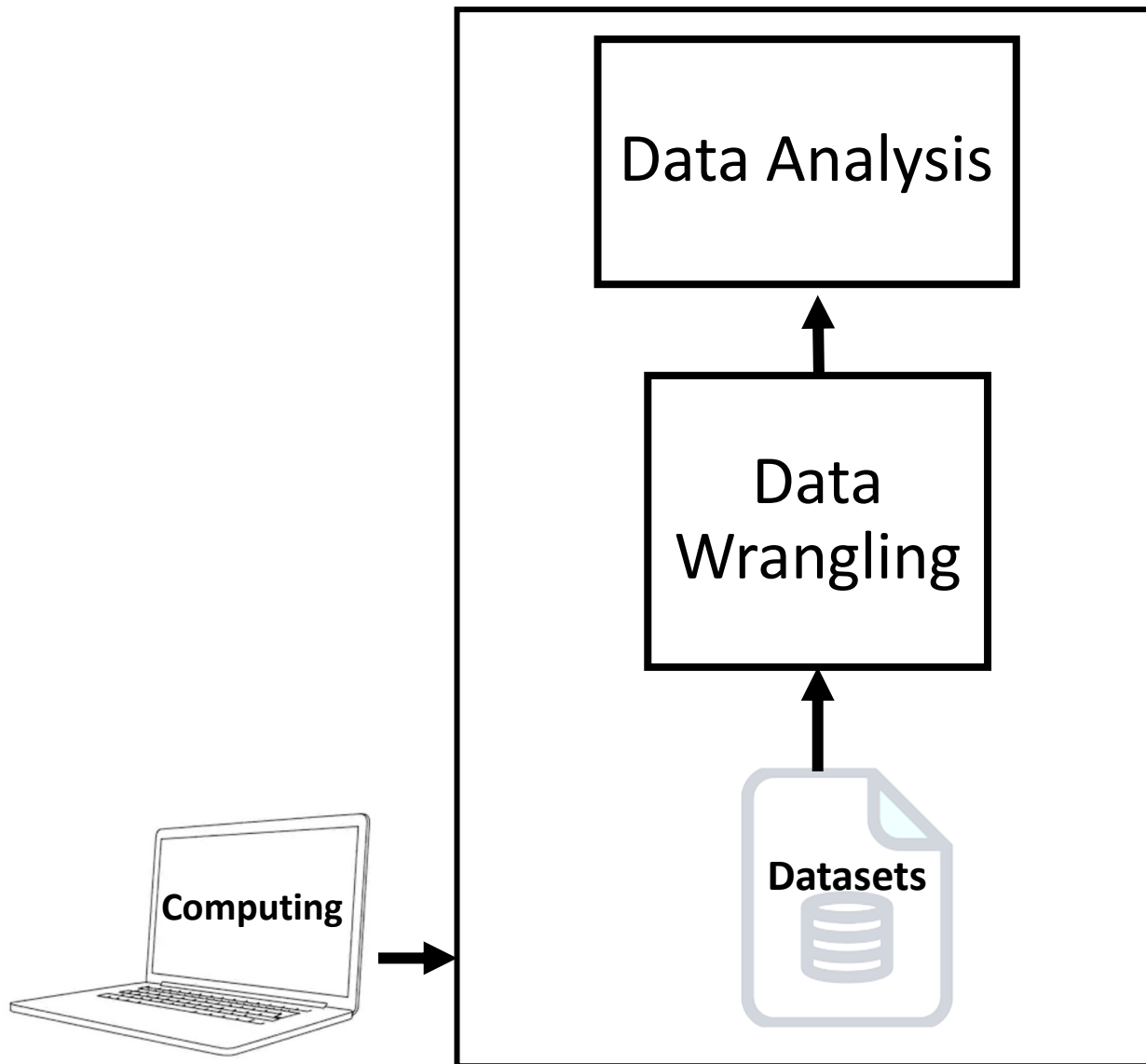
## Some Other FAIR Principles

- Findable** – can the user find in things at the helicopter level
- Accessible** – can the use get to the information in the weeds
- Informative** – does it tell the user what they need to know
- Robust** – is the info stable (e.g. 3 waves later on)



# Data Access

- End User Licenses have been very effective
- Increasing use of Special License (e.g. Occupational SOC codes) is this really necessary?





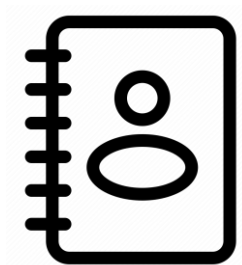
# Data Access

- Secure lab work is more time consuming than desktop work
- Is too much linked admin data unnecessarily held within the secure lab?

# Linking to Administrative Data Sources

A characteristic of micro-level administrative social science data sets is that they usually have a large number of observations ( $n$ ), for example individuals, but a smaller number of social science related explanatory variables ( $k$ ) than would be the case for social surveys





**An investigation of the consistency of GCSE qualifications data in administrative educational records and a national social survey**

Sarah Stopforth, University of York

Roxanne Connelly, University of Edinburgh

Vernon Gayle, University of Edinburgh

- We advise researchers to access the linked GCSE data from the National Pupil Database (NPD)
- Beware the raw NPD data requires a large amount of data wrangling to prepare – measures are unsuitable for immediate use in analyses (e.g. in the pupil-level dataset)

Is the quality of the admin data = to the quality of the survey data?

Is the quality of documentation of the admin data = to the quality of the survey documentation?

# Do We Need A Scottish Strategy?

*Radical Statistics      Issue 97*

## **Scottish Social Survey Data, Past, Present and Future – Does Scotland Need its Own Data Strategy?**

*Vernon Gayle, Christopher Playford  
and Paul Lambert*

Gayle, V., Playford, C. and Lambert, P., 2008. Scottish Social Survey Data, Past, Present and Future—Does Scotland Need its Own Data Strategy?. *Radical Statistics*, 97, pp.82-97.