Data Wrangling: Organising and Enabling Data ©

Dr Diarmuid McDonnell & Prof. Vernon Gayle
AQMEN
University of Edinburgh

March 2019

Welcome

Why are we here today?

U.K. Industrial Strategy aims to utilise big data to improve economic performance and increase productivity.

Major barrier is the lack of a suitably trained workforce.

U.K. social science stakeholders (e.g. ESRC, Nuffield) believe this discipline can make a major contribution to Industrial Strategy.

Why this type of training?

"Many organizations can barely find a way to use their R/Python programmers on reasonable datasets."

"The piece missing from the data science movement right now is really simple: intelligent application of data science tools."

https://www.linkedin.com/pulse/data-science-dead-5-years-less-justin-b-dickerson-phd-mba-pstat- accessed 16.07.2018.


Administrative Data Research Centre Scotland
An ESRC Data Investment
Better Knowledge Better Society
Supporting researchers to use administrative data for economic and social research
Enabling studies that benefit the public
Helping inform and improve policy-making
W. adrn.ac.uk
E. adrc-s@ed.ac.uk
AQMeN
Applied Quantitative Methods Network
In partnership with the Fair Institute of Scotland and the Scottish Government

What is the social science contribution?

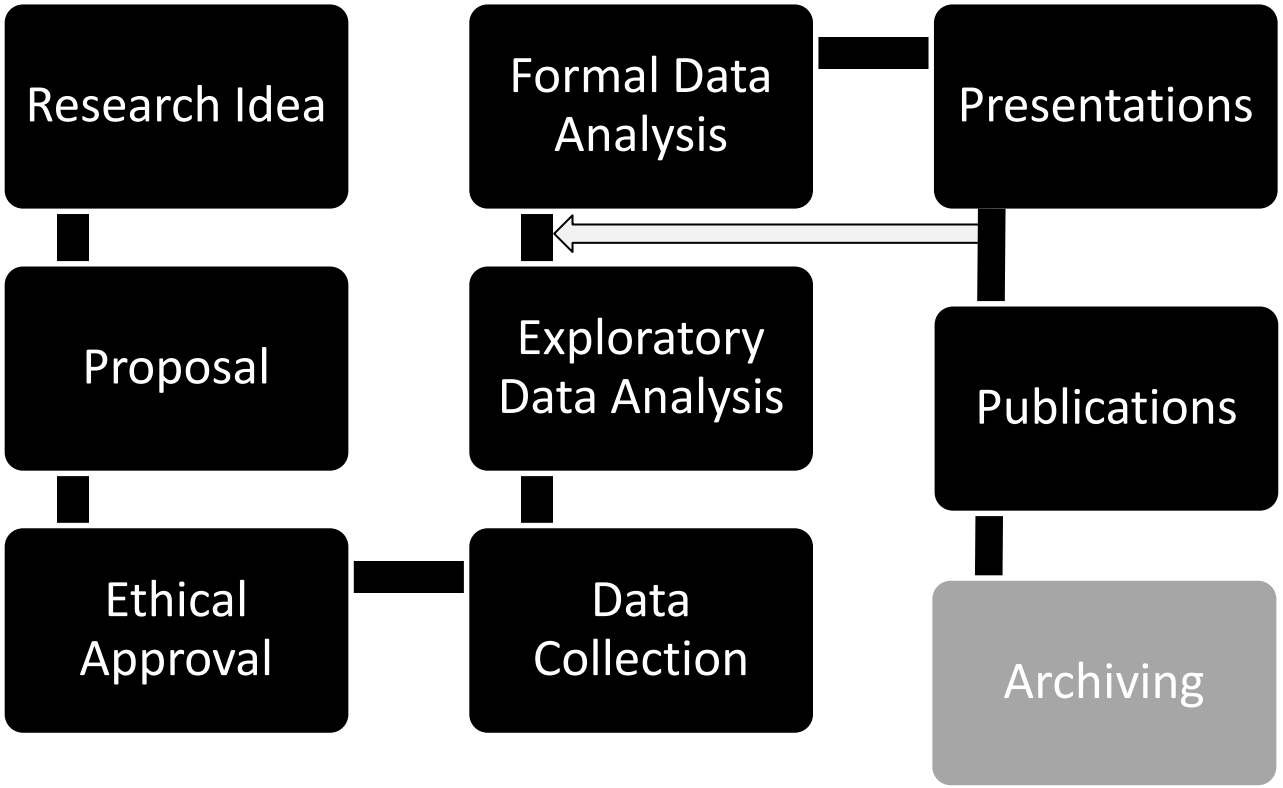
Data Science ≠ Computer Science

Big Data ≈ Small Data

QM Social Scientists = Data Literate

Data Wrangling

The 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?

Be honest…

Have you ever struggled to identify which data file is the correct one?

chapter1_2014.dat

chap1_2014.dat


383
791

The Workflow

- Planning, organising and documenting work
- This includes…

Cleaning data

Analysing data

Presenting results

Backing up and archiving material

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

Better supporting YOU and what YOU DO

Not changing you into something YOU ARE NOT

Shopping without a list?

Cooking with a list?

APPROVED B-17F and G CHECKLIST

REVISED 3-1-44

PILOT'S DUTIES IN RED

COPILOT'S DUTIES IN BLACK

BEFORE STARTING

1. Pilot's Preflight—COMPLETE
2. Form 1A—CHECKED
3. Controls and Seats—CHECKED
4. Fuel Transfer Valves & Switch—OFF
5. Intercoolers—Cold
6. Gyros—UNCAGED
7. Fuel Shut-off Switches—OPEN
8. Gear Switch—NEUTRAL
9. Cowl Flaps—Open Right—OPEN LEFT—Locked
10. Turbos—OFF
11. Idle cut-off—CHECKED
12. Throttles—CLOSED
13. High RPM—CHECKED
14. Autopilot—OFF
15. De-icers and Anti-icers, Wing and Prop—OFF
16. Cabin Heat—OFF
17. Generators—OFF

STARTING ENGINES

1. Fire Guard and Call Clear—LEFT Right
2. Master Switch—ON
3. Battery switches and inverters—ON & CHECKED
4. Parking Brakes—Hydraulic Check—On-CHECKED
5. Booster Pumps—Pressure—ON & CHECKED
6. Carburetor Filters—Open
7. Fuel Quantity—Gallons per tank
8. Start Engines: both magnetos on after one revolution
9. Flight Indicator & Vacuum Pressures CHECKED
10. Radio—On
11. Check Instruments—CHECKED
12. Crew Report
13. Radio Call & Altimeter—SET

ENGINE RUN-UP

1. Brakes—Locked
2. Trim Tabs—SET
3. Exercise Turbos and Props
4. Check Generators—CHECKED & OFF
5. Run up Engines

BEFORE TAKEOFF

1. Tailwheel—Locked
2. Gyro—Set
3. Generators—ON

AFTER TAKEOFF

1. Wheel—PILOT'S SIGNAL
2. Power Reduction
3. Cowl Flaps
4. Wheel Check—OK right—OK LEFT

BEFORE LANDING

1. Radio Call, Altimeter—SET
2. Crew Positions—OK
3. Autopilot—OFF
4. Booster Pumps—On
5. Mixture Controls—AUTO-RICH
6. Intercooler—Set
7. Carburetor Filters—Open
8. Wing De-icers—Off
9. Landing Gear
   a. Visual—Down Right—DOWN LEFT Tailwheel Down, Antenna in, Ball Turret Checked
   b. Light—OK
   c. Switch Off—Neutral
10. Hydraulic Pressure—OK Valve closed
11. RPM 2100—Set
12. Turbos—Set
13. Flaps ⅓—⅓ Down

FINAL APPROACH

14. Flaps—PILOT'S SIGNAL
15. RPM 2200—PILOT'S SIGNAL

In the late 1930s, military aviators in the American Army and Navy began using aviation checklists. Checklist became part of a new paradigm for how to fly, which consisted of

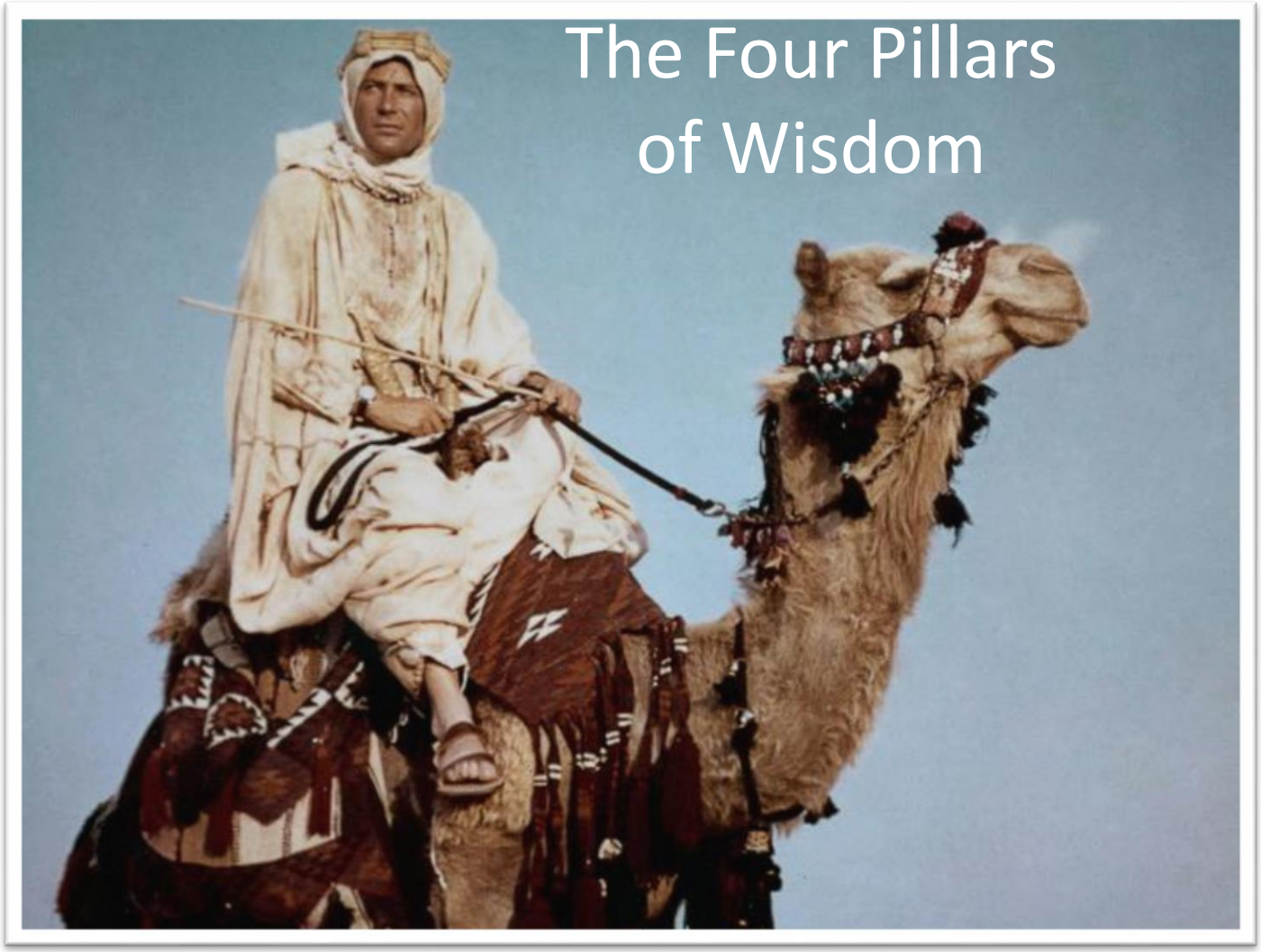
- Elaborate standardized procedures for many activities
- Checklists to ensure all critical steps had been done
- Quantitative tables and formulas that specified the best settings, under different conditions, for speed, engine RPM, gasoline/air mixture, engine cooling, and many other parameters.

This new paradigm (Standard Procedure Flying) had a major influence on reducing aviation accidents and increasing military effectiveness during World War II, particularly because of the rapidly increasing complexity of military aircraft, and the huge number of new pilots.

Despite the benefits of Standard Procedure Flying for both safety and efficiency, by the end of WWII only a few air forces had fully embraced it

Roger Bohn http://www.vs29.org/Links/NATOPS/SOP-bohn-2013-1.pdf


The Platinum Standard
tt.mop.com


The Four Pillars
of Wisdom

A Planned Workflow Has Benefits

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

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

The best habit that you can get into

is to get into good habits!

The Workflow

Drukker's dictum: Never type anything that you can obtain from a saved result

My dictum (Gayle's dictum):

You can't be too fit or have too many publications

However…

- 500+ scientific publications in peer reviewed journals (15,000+ citations and an H-index of 66)
- Has run more than 70 marathon and ultra-marathon races, including seven 90km Comrades Marathons and fifteen 56km Two Oceans Marathons

http://www.essm.uct.ac.za/ESSM/Tim_Noakes

- Over 20 Ultra Marathons including the Western States 100 mile race
- 1480 citations since 2011

https://www.stat.berkeley.edu/~stark/index.html

Long's Law

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

Corollary 1:

Nobody likes to write documentation

Corollary 2:

Nobody every regrets having written documentation

Long's Law

Has anyone in the history of data analysis ever said

"these files are too well documented"

scarc.library.oregonstate.edu/coll/pauling/rnb/index.html

As with many scientists, Linus Pauling utilized bound notebooks to keep track of the details of his research as it unfolded. A testament to the remarkable length and diversity of Dr. Pauling's career, the Pauling Papers holdings include forty-six research notebooks spanning the years of 1922 to 1994 and covering any number of the scientific fields in which Dr. Pauling involved himself. In this regard, the notebooks contain many of Pauling's laboratory calculations and experimental data, as well as scientific conclusions, ideas for further research and numerous autobiographical musings.

Research Notebook 01
1922
Research Notebook 02
1922-1923, 1932, 1934, 1936, 1973, 1985
Research Notebook 03
1923-1925
Research Notebook 04
1923-1924, 1928-1930
Research Notebook 05

Research Notebook 13
1935-1936, 1938-1939
Research Notebook 14
1936-1939, 1949, 1952
Research Notebook 15
1935, 1937, 1968
Research Notebook 16
1935-1956
Research Notebook 17
1939-1941, 1971, 1988

Research Notebook 24
1953, 1956, 1962, 1963, 1967, 1968, 1969, 1970, 1973
Research Notebook 25
1958, 1964-1966
Research Notebook 26
1955, 1964-1969, 1974-1976, 1980-1982, 1987, 1990-1991
Research Notebook 27
1952-1954, 1960-1961, 1964, 1971-

Research Notebook 35b
1938-1939, 1946, 1955, 1968, 1986-1988
Research Notebook 36
1980-1981, 1986-1987
Research Notebook 37
1971, 1983
Research Notebook 38
1980-1981, 1983, 1985, 1989
Research Notebook 39

2 June 1973
Portola Valley, Cal.
Linus Pauling.

Golden Wedding Anniversary

Three days ago Ava Helen and I celebrated our 50th wedding anniversary. We had been married in Salem, Oregon, on 17 June 1923.

Our celebration was at the ranch. Our four children were there, also Linda's husband (Barclay) and their four boys; also Art and Laureke Robinson; also Art Cherkin; also Ewan Cameron from Scotland, his wife, and two children; also L. Margo Miller, Ava Helen's sister; also my two sisters, Pauline Pauling-Wey and Lucille Jenkins

- Improving the workflow with a modest amount of effort

- The less experience you have the better
  - start from the very beginning

ALL SERIOUS WORK MUST BE REPRODUCIBLE!

There MUST be an audit trail

Why is it all so difficult?

Social science data tends to come in messy formats

Administrative data often is even more complex in nature than social survey data

Minor decisions have major consequences…

Which cases?

Which variables?

How to code (e.g. education)?

How to recode?

Where do I truncate?

Minor decisions have major consequences…

Which cases?

Which variables?

How to code (e.g. education)?

How to recode?

Where do I truncate?

Can I trace these decisions in my audit trail?

template* ×

```
STOP

/**

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

Next Actions:



Author:

Project:

Sub-project:

Date of Next Meeting (or supervision):

Latest Update:

Previous Updates:


Useful information:
http://www.samaritans.org/ (08457 90 90 90)

*********************************************************************************
```

File Naming Protocols

File Name =

name_date_depositor's initials_version_type

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)

Why is it all so difficult?

Poor discipline and insufficient documentation

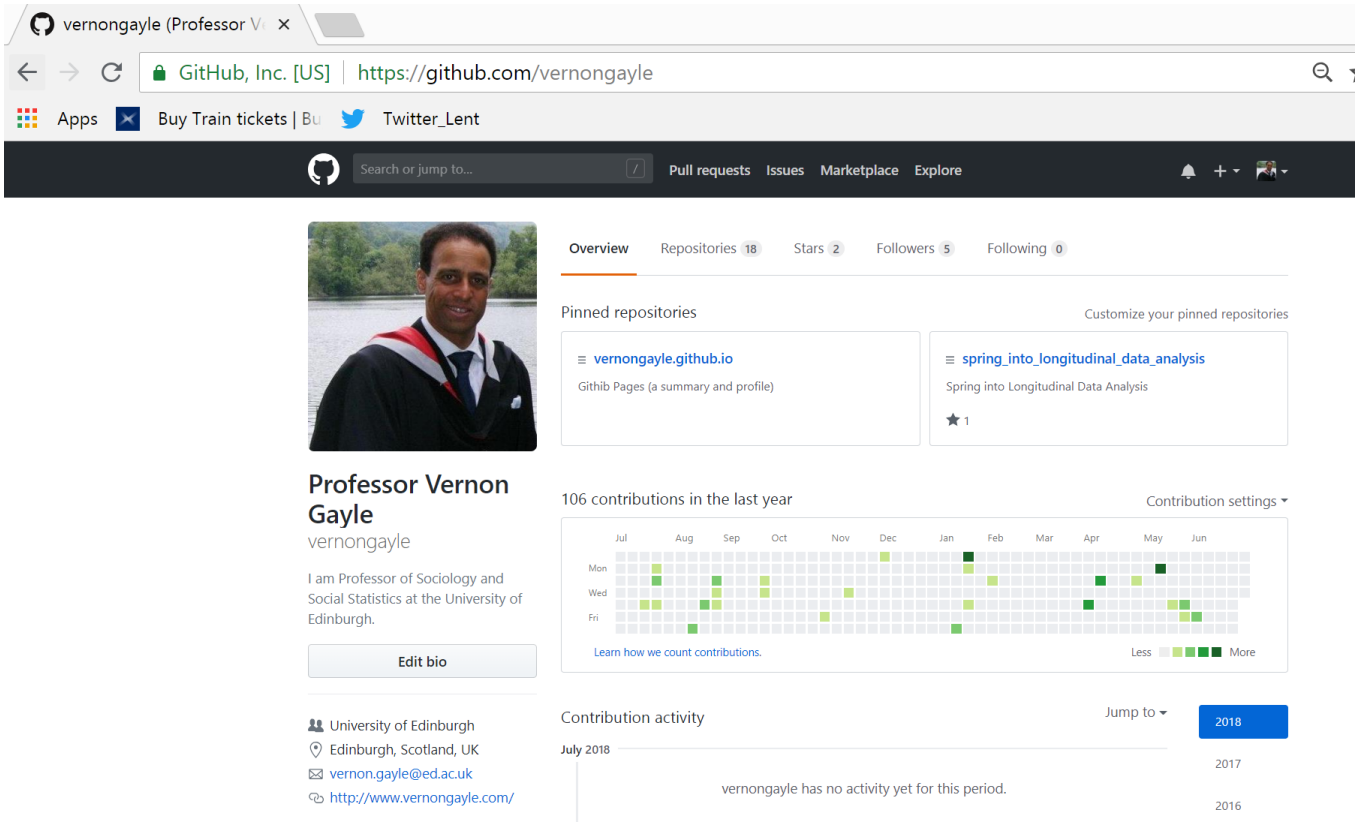
jupyter

Install About Resources Documentation NBViewer Widgets Blog Donate

Juila, Python and R almost spell JuPyteR

Open source, interactive data science and scientific computing across over 40 programming languages.

https://jupyter.org/


vernongayle (Professor V
GitHub, Inc. [US] | https://github.com/vernongayle
Apps
Buy Train tickets | Bu
Twitter_Lent
Search or jump to...
Pull requests
Issues
Marketplace
Explore
Overview
Repositories 18
Stars 2
Followers 5
Following 0
Pinned repositories
Customize your pinned repositories
vernongayle.github.io
Githib Pages (a summary and profile)
spring_into_longitudinal_data_analysis
Spring into Longitudinal Data Analysis
1
Professor Vernon Gayle
vernongayle
I am Professor of Sociology and Social Statistics at the University of Edinburgh.
Edit bio
University of Edinburgh
Edinburgh, Scotland, UK
vernon.gayle@ed.ac.uk
http://www.vernongayle.com/
106 contributions in the last year
Contribution settings
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Mon
Wed
Fri
Learn how we count contributions.
Less
More
Contribution activity
Jump to
2018
2017
2016
July 2018
vernongayle has no activity yet for this period.


Google docs


Dropbox


box


OneDrive


EVERNOTE


GitLab


asana


Bitbucket

Workshop

R

- Growing in popularity (e.g. data science, statistics, science etc.)
- Popular with statisticians
- Free (open source)
- Difficult to learn
- Development and support is not commercial
- Help resources are under-developed

Programme

Mix of talks and self-directed practical activities.

Data Wrangling challenge (“Hackathon”).

Tutor and peer support.

Use of a variety of data sets, especially messy administrative records.

Top tips

1. Ask plenty of questions.
2. Take your time.
3. Complete as many of the tasks and exercises, and answer as many of the questions as you can.
4. Annotate your work.
5. Be positive.

Estimating Work Time…

Good Luck

Our aim is to equip you, as rapidly and painlessly as possible, with a proficiency in data wrangling using R.

We think it is an ambitious yet achievable goal.

Them: “Are you any good at data wrangling?”

You: “Yes, yes I am.”