Jupyter Notebooks

Office of Planning & Analysis
UC Berkeley
March 2020

Professor Vernon Gayle

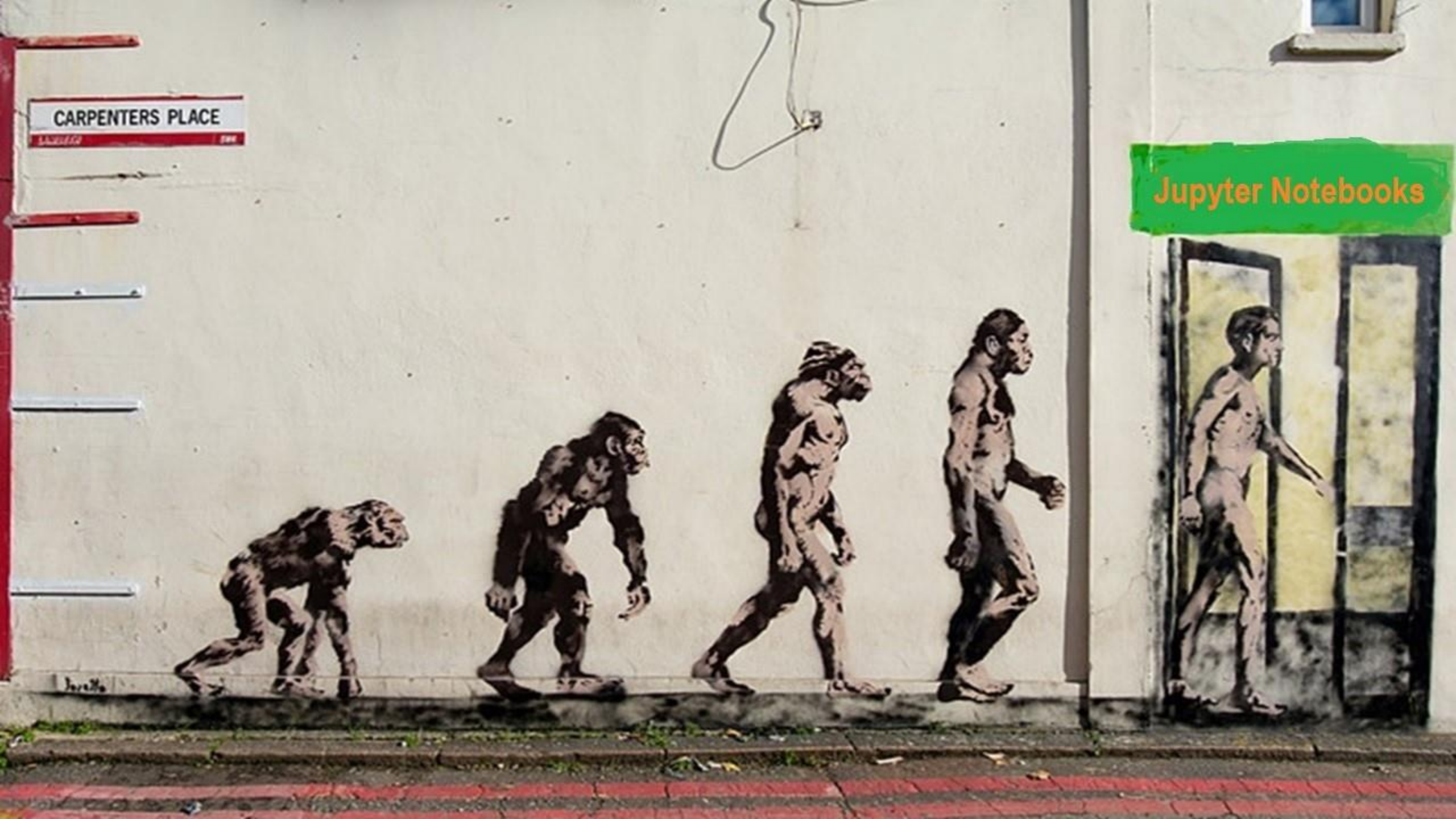
vernon.gayle@ed.ac.uk
@Profbigvern
https://github.com/vernongayle

NCRM National Centre for Research Methods

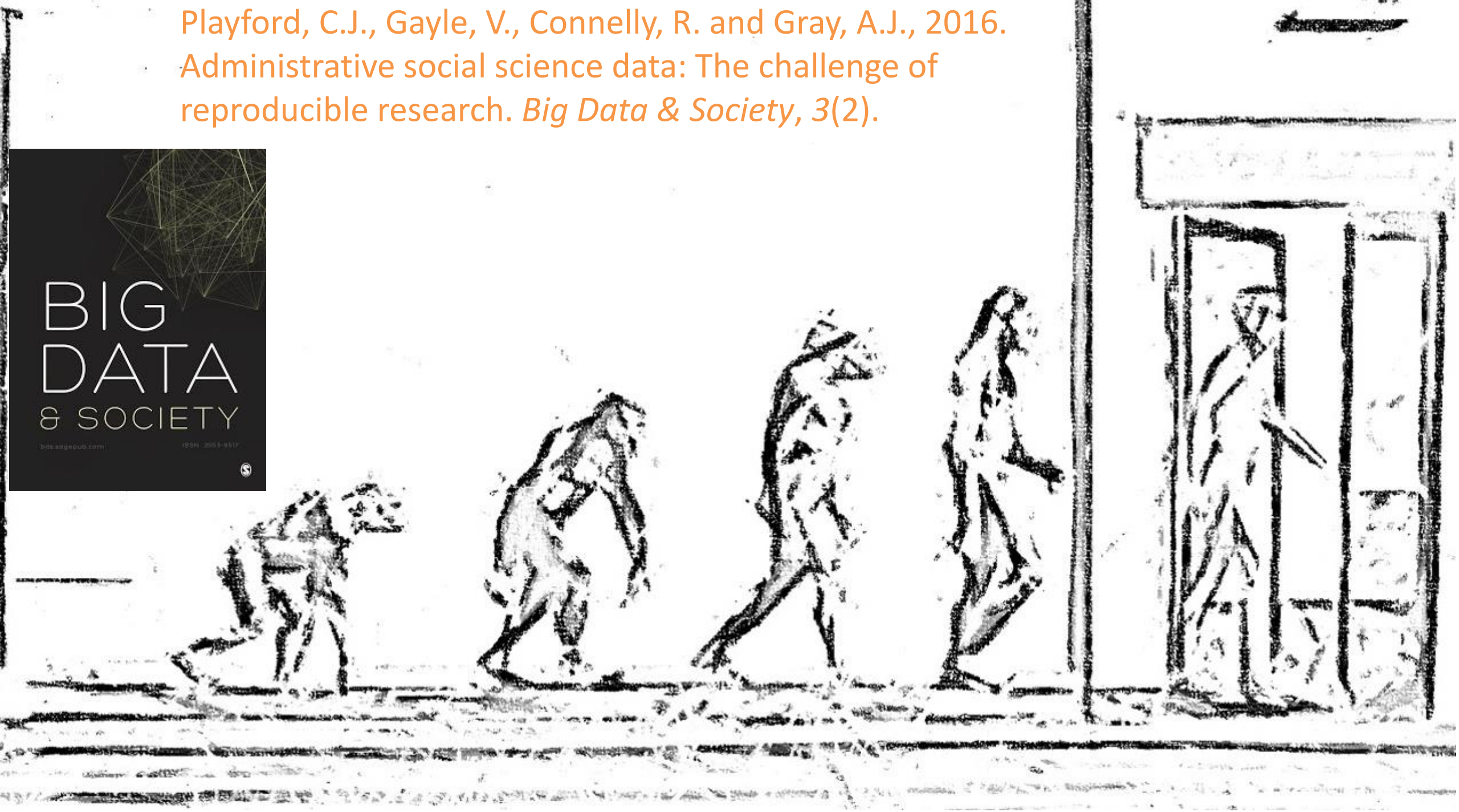
Research Serendipity

Some ultramarathons I've run:

Event	Distance	Year	State	Climb
Angel Island 50k	50 km	2005	CA	4,200'
Backward Western States	100+ mi	2004	CA	21,970'
Backyard Hundred	100 mi	2006	CA	~20,000'
The Bear	100.33 mi	2004	ID	21,061'
		8	CA	~15,000'
		5	WY	18,308'
			CA	3,890'
			WA	20,470'
			CA	28,102'
			CA	7,800'
			CA	~10,000'
			CA	8,200'
			CA	18,552'
			CA	15,600'
			Z	~6000'
			K	
				3,710'
				4,750'
				11,300'
				11,300'


CARPENTERS PLACE
Jupyter Notebooks

Playford, C.J., Gayle, V., Connelly, R. and Gray, A.J., 2016. Administrative social science data: The challenge of reproducible research. *Big Data & Society*, *3*(2).



The Case for Greater Transparency

1. Increase the capacity to understand how the research was conducted
2. Help others evaluate the analyses undertaken
3. Aid the detection of errors and inconsistencies
4. Facilitate the incremental development of work
5. Contribute to limiting negative research practices
6. Provide extra safeguards against nefarious practices
7. Improve confidence in results

What is the Problem?

- Empirical results cannot be reproduced because of a lack of transparency in the research process (Baker, 2016)

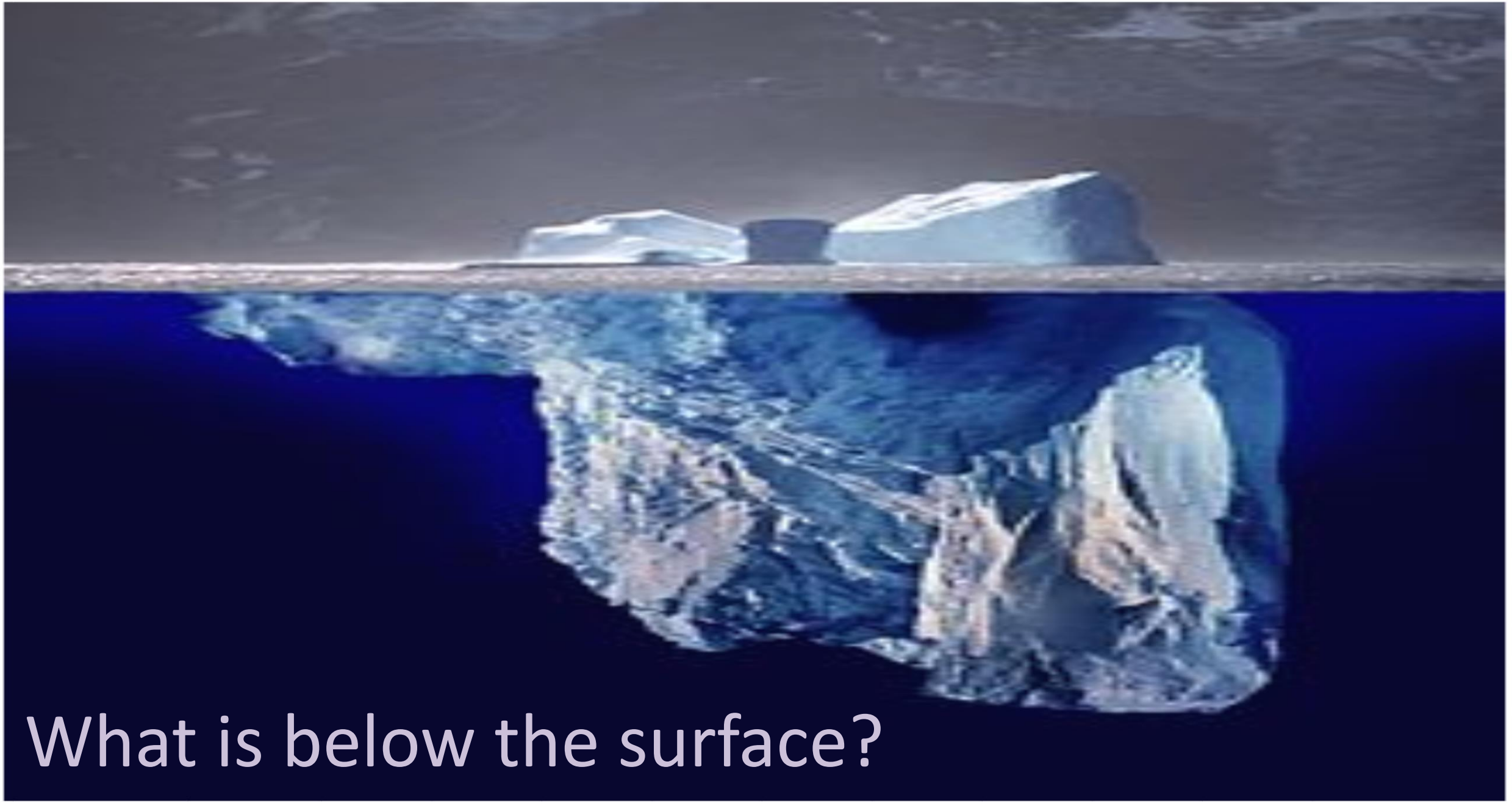
- Impossible to verify the results presented in many research papers, reports etc. (Christensen, Freese, & Miguel, 2019)

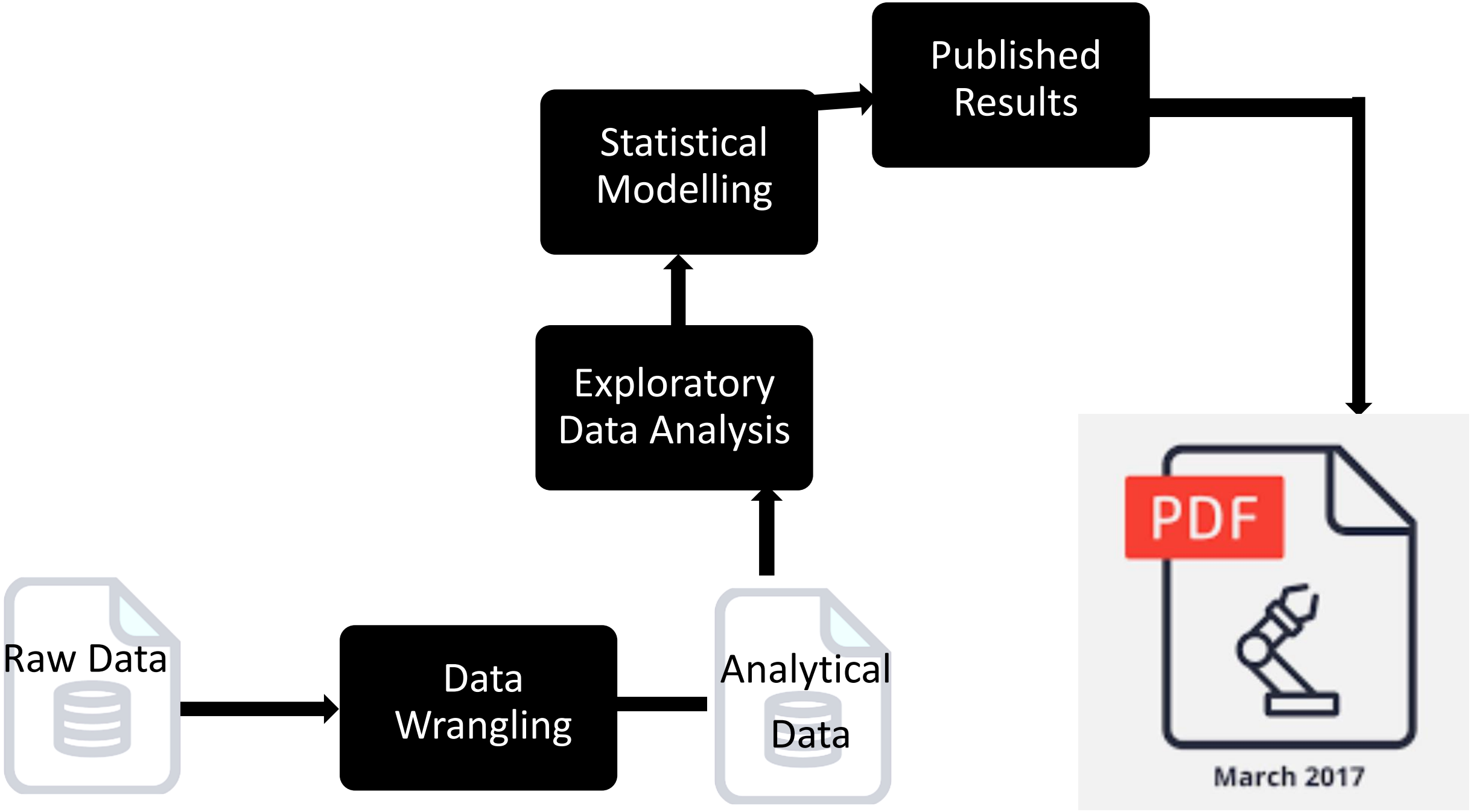
Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. doi: 10.1038/533452a

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. Oakland, California: University of California Press.

Can we **reverse-engineer** results?

The final publication if the ‘tip of the iceberg’

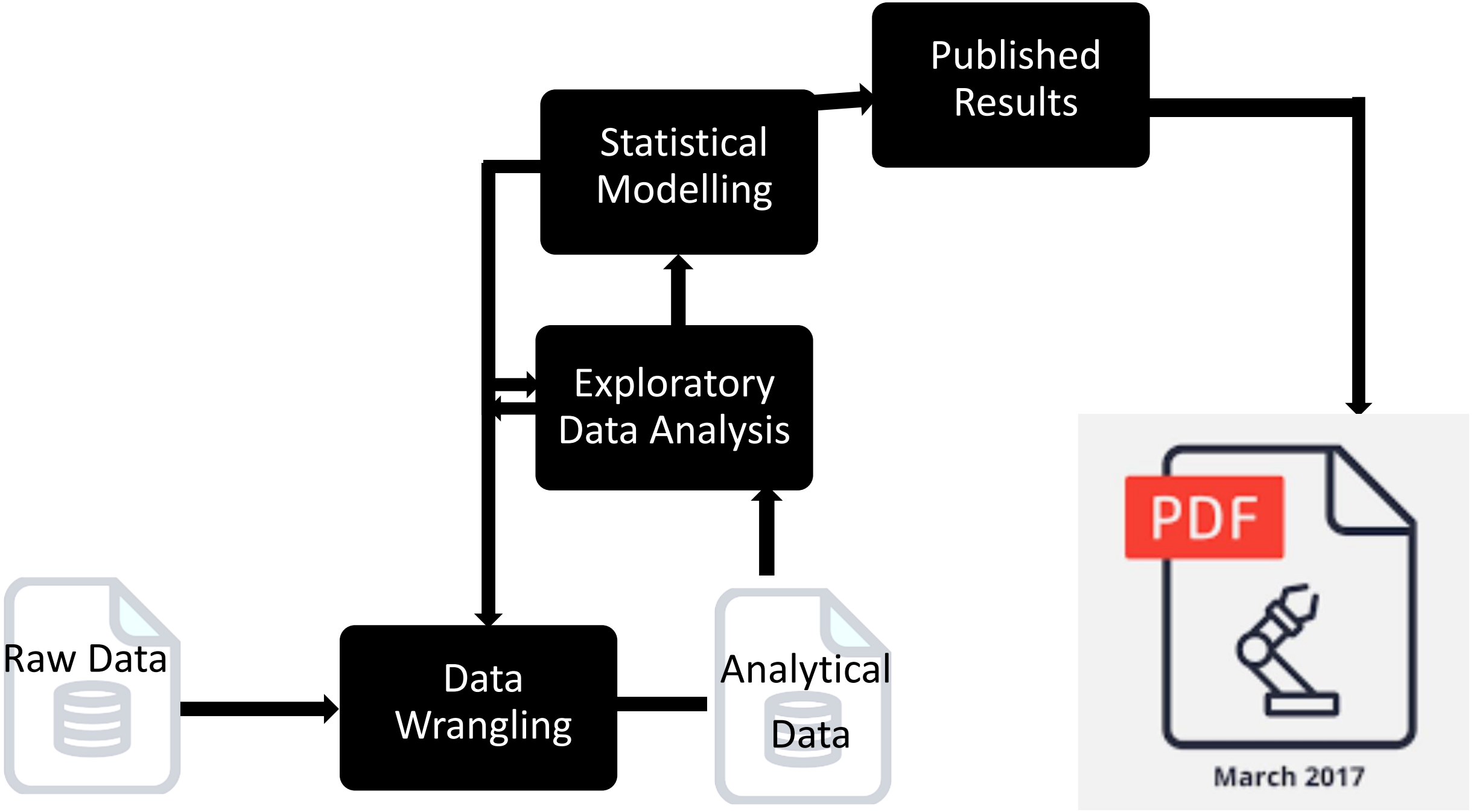

What is below the surface?


Raw Data
Data Wrangling
Analytical Data
Exploratory Data Analysis
Statistical Modelling
Published Results
PDF
March 2017


‘Raw’
Data Download
Data Wrangling
Analytical
Data

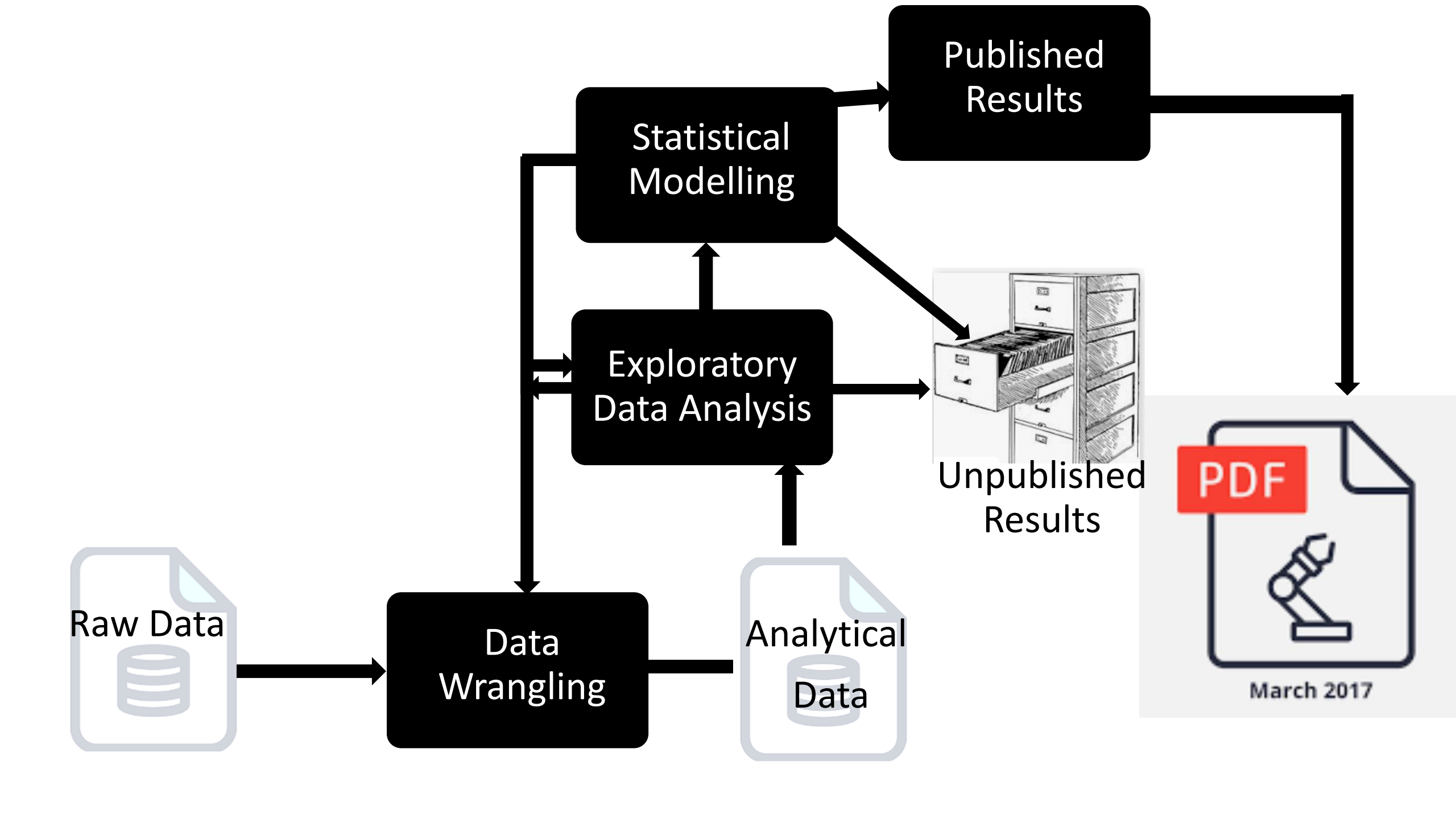
Data Wrangling (Black Box)

- 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

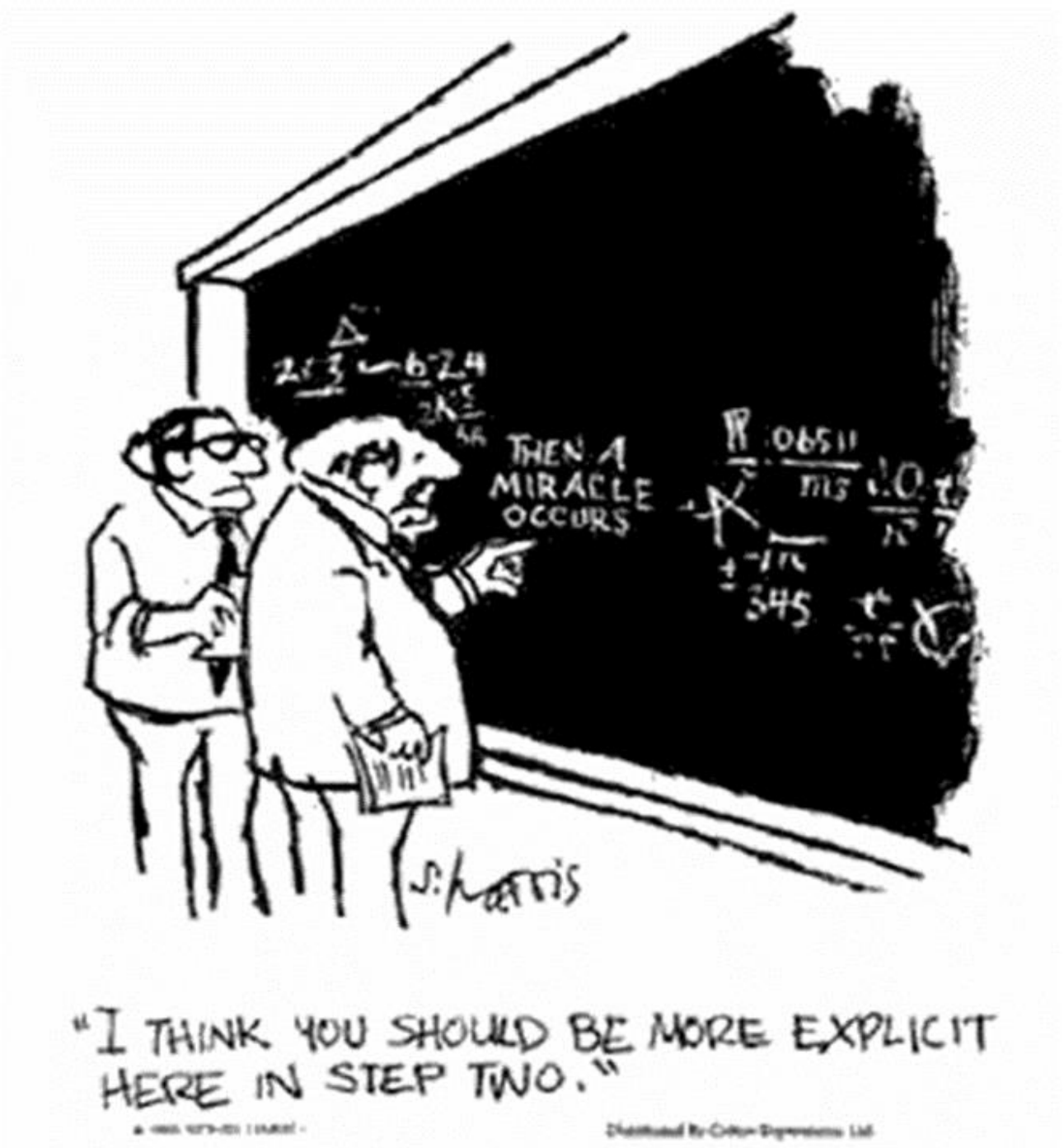

Published Results
Statistical Modelling
Exploratory Data Analysis
Raw Data
Data Wrangling
Analytical Data
PDF
March 2017

Data Analysis (Black Box)

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


Published Results
Statistical Modelling
Exploratory Data Analysis
Unpublished Results
PDF
March 2017
Raw Data
Data Wrangling
Analytical Data

‘Raw’
Data Download



PDF

March 2017

Transparent and Reproducible

*Make public sufficient information so that a third party who is unconnected with the original research can '**duplicate**' the work and obtain the same results (without contacting the original authors)*

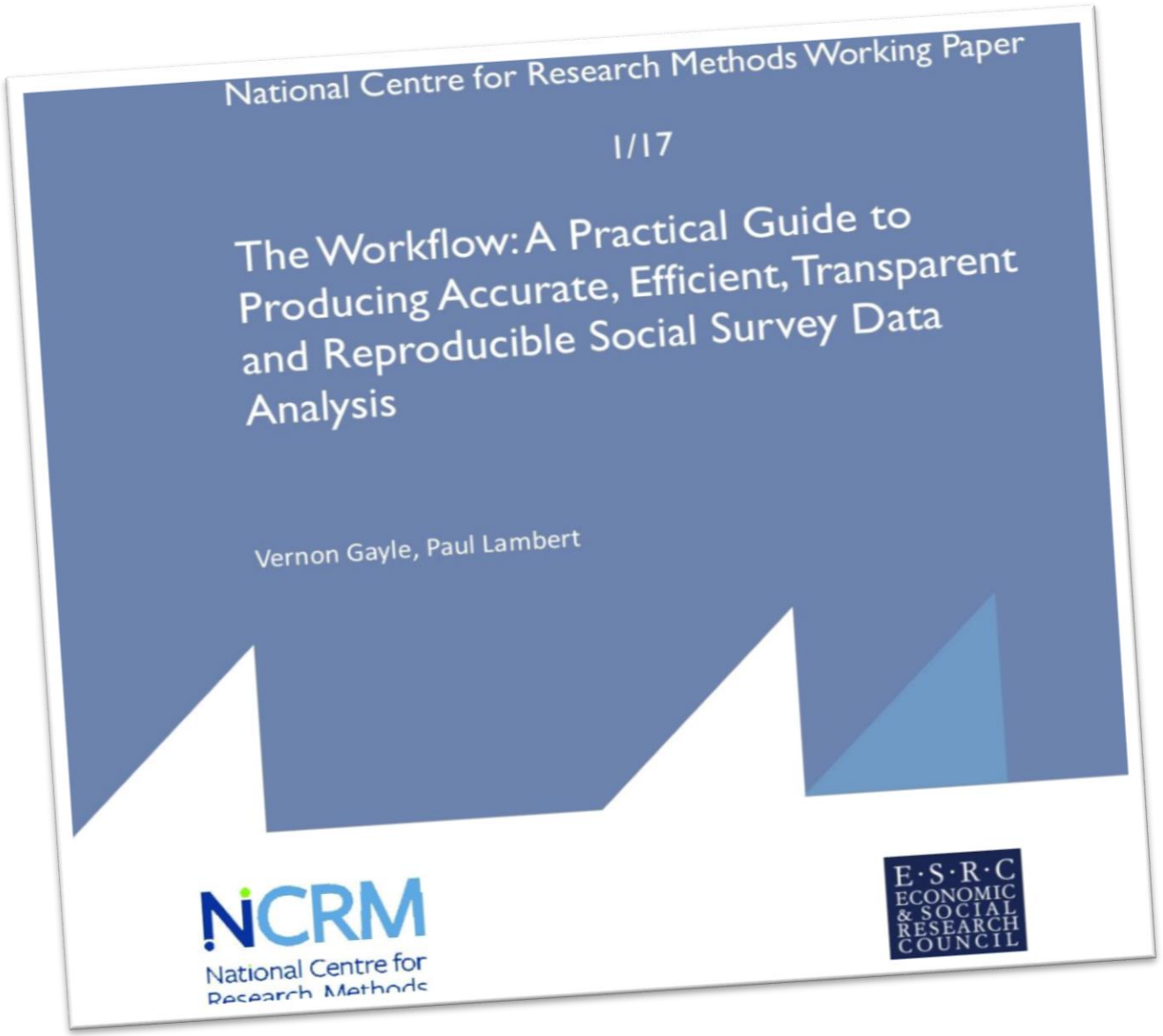
Replication
(extends the original work)

1. Additional measures
2. Alternative measures
3. New data
4. Different statistical analytical techniques

Or any combination of these four components

Make the **COMPLETE** workflow public

Raw Data



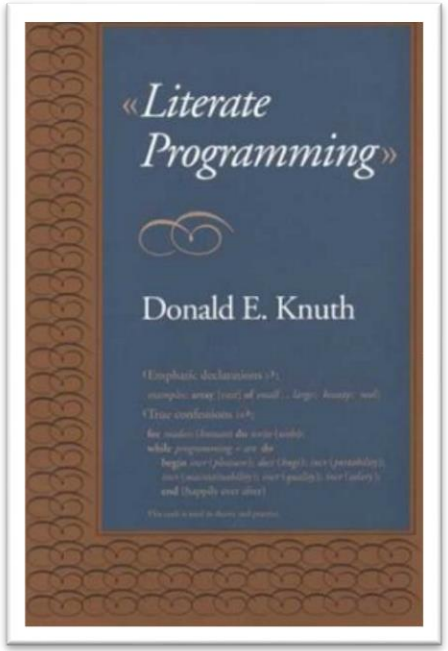
http://eprints.ncrm.ac.uk/4000/

Literate Computing

Fernando Perez says

Literate Computing is the weaving of a narrative directly into a live computation, interleaving text with code and results to construct a complete piece that relies equally on the textual explanations and the computational components, for the goals of communicating results in scientific computing and data analysis.

http://blog.fperez.org/

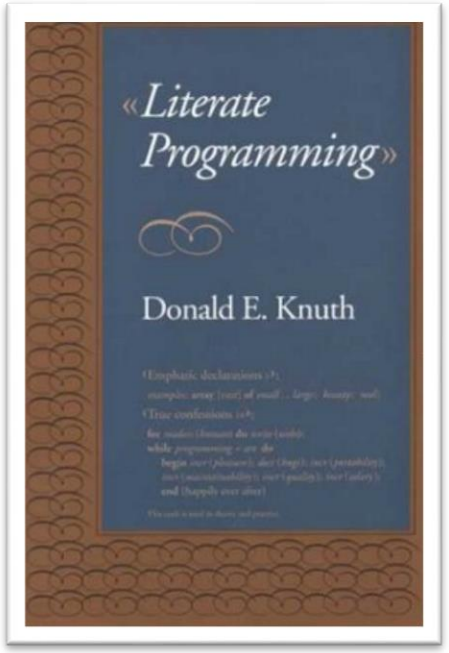


Knuth says

Treat your program as literature

People publish scores of symphonies they don’t just listen to them

http://tinyurl.com/he5aagf



Knuth says

Treat your program as literature

People publish scores of symphonies they don't just listen to them

Both people and computers should be able to read your program

If others can read my program I will understand my own program better

http://tinyurl.com/he5aagf

Advice from Computer Science

Write a book with the executable code in it

It will be big!

It does not matter how big your book is…

> *“as long as you have a table of contents (for the big topics) and an index for the small things”*

Timothy Daly - Computer Scientist

Example of a ‘literate’ comment

The occupational information we are going to use for our parental social class measure comes from the new occupational coding files (SN7023).

Gregg, P. (2012). Occupational Coding for the National Child Development Study (1969, 1991-2008) and the 1970 British Cohort Study (1980, 2000-2008). [data collection]. University of London. Institute of Education. Centre for Longitudinal Studies, [original data producer(s)]. UK Data Service. SN: 7023.

"Researchers from the Avon Longitudinal Study of Parents and Children (ALSPAC), based at the University of Bristol, worked on data from selected waves of the NCDS and BCS70. To create occupational code classifications, the computerised questionnaire response text strings were converted into comma separated value (CSV) files and processed using the CASCOT (Computer Assisted Structured COding Tool) software programme, which used automatic and semi-automatic processing to assign Standard Occupational Classification 2000 (SOC2000) codes (SOC2000) to entries."

The NS-SEC Full Version is recoded to the 8 category version using the NS-SEC documentation available here. See the classes and collapses of NS-SEC here.

In [26]:
```
use $path1\ARCHIVE\NCDSBCS_OCCS\ncds2_occupation_coding_father.dta, clear

describe
```

Example of a ‘literate’ structure

Table of Contents:

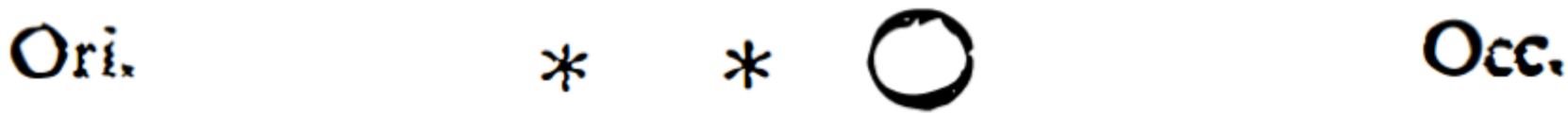
What Can We Learn From Other Research Areas?

Research Objects

Research Objects enable the identification and exchange of scholarly information

Primary goal is to associate related resources in a scientific investigation

But on January 10th the stars appeared in the following position with regard to Jupiter; there were two only, and both on the east side



of Jupiter, the third, as I thought, being hidden by the planet.

Sidereus Nuncius (Galilei 1610)


Altair
Mars
Jupiter
Saturn
δ-Centaurids
ν-Normids
E
SE
S
SW
Earth, Berkeley, 0 m
FOV 116°
18.1 FPS
2020-03-10 05:32:10 UTC-07:00

jupyter

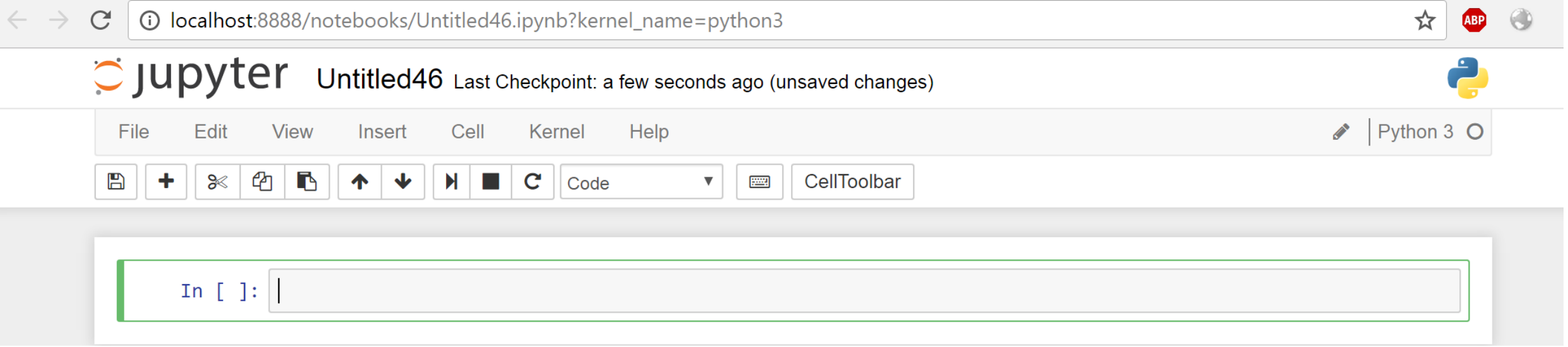
Install About Resources Documentation NBViewer Widgets Blog Donate

Juila, Python and R almost spell JuPyteR

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

https://jupyter.org/

- Easy documentation alongside research code
- ‘Language agnostic’ 40+ languages
- Rich visual outputs
- Big data tools e.g. python
- Teaching and training
- Collaborative work
- Portability (publication) easy to share



Open source web application

Creates documents which include live code, output and explanatory text

Single platform for the complete workflow

Code

In [4]:
```
summarize
```

Output

In [4]:
```
summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
case	1,580	517.7411	284.8605	1	1003
femp	1,580	.6455696	.4784918	0	1
mune	1,580	.0740506	.2619362	0	1
time	1,580	7.2	3.981019	0	13
und1	1,580	.0746835	.2629633	0	1
und5	1,580	.2974684	.4572891	0	1
age	1,580	36.01013	9.114841	18	60

Text (Markdown)

In [4]:
```
summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
case	1,580	517.7411	284.8605	1	1003
femp	1,580	.6455696	.4784918	0	1
mune	1,580	.0740506	.2619362	0	1
time	1,580	7.2	3.981019	0	13
und1	1,580	.0746835	.2629633	0	1
und5	1,580	.2974684	.4572891	0	1
age	1,580	36.01013	9.114841	18	60

The data mirror a real example of data analysed in Davies et al. (1992).

The dataset is a panel of 155 married women.

Davies, Richard B., Peter Elias, and Roger Penn. "The relationship between a husband's unemployment and his wife's participation in the labour force." Oxford Bulletin of Economics and Statistics 54.2 (1992): 145-171.

Markdown

- *Markdown* is an easy way to write documents
- It is written in what computer geeks like to call 'plaintext'
- Plaintext is just the regular alphabet plus a few other familiar symbols (for example the asterisk *)
- Unlike cumbersome word processing applications, text written in Markdown can be easily shared between computers

Markdown

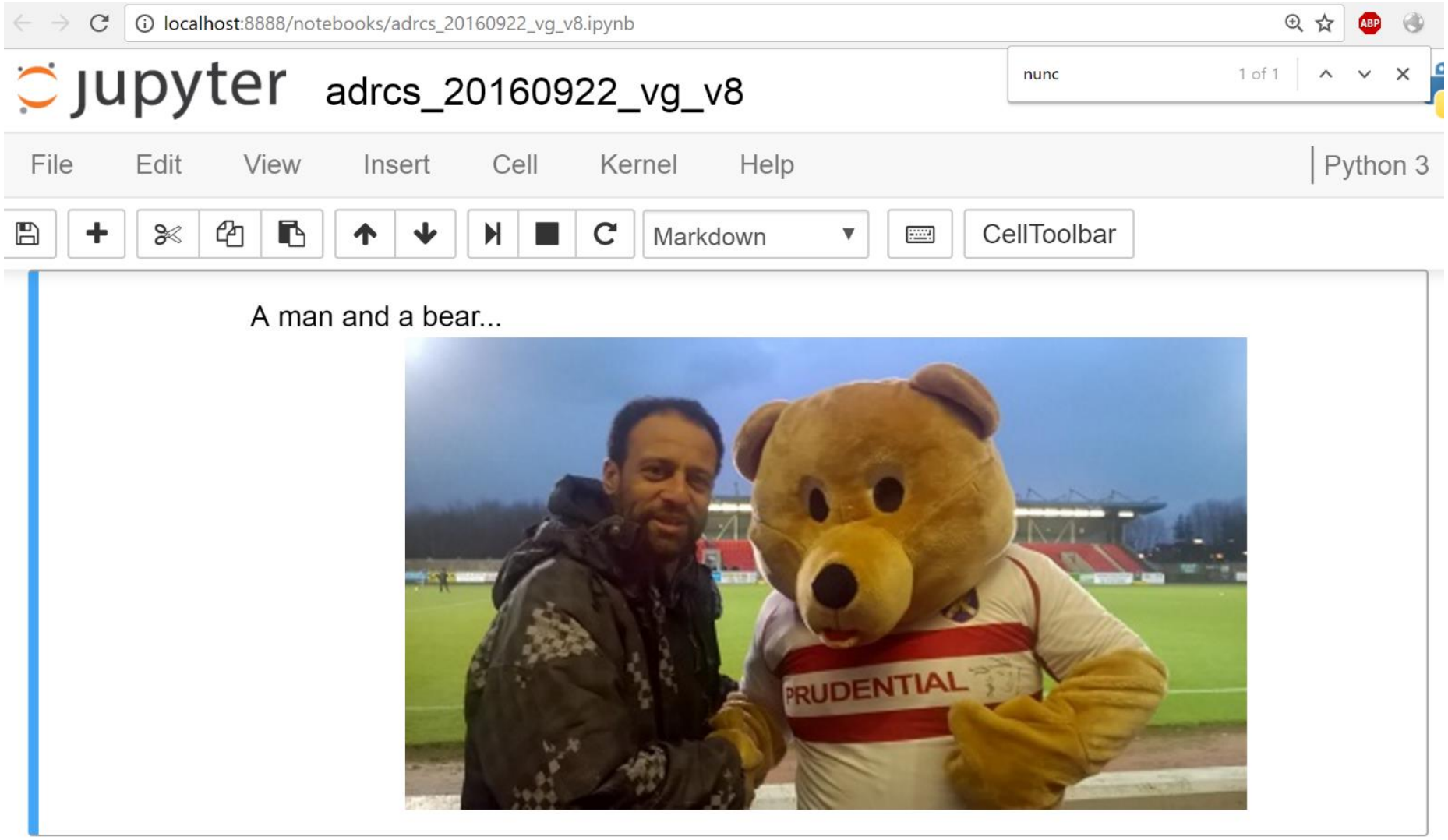
- It’s quickly becoming the writing standard in some academic areas and in science

- Websites like GitHub and reddit use Markdown to style their comments

- Here is a summary of *Markdown* codes https://en.wikipedia.org/wiki/Markdown#Example

- If you have half an hour you can learn *Markdown* here http://www.markdowntutorial.com/ (try a different browser)

Images within the notebook cell…



LaTeX

«Lah-tech» rhymes with «Bertolt Brecht»

to render cell contents as LaTeX

In [8]:
```
%%latex
\begin{align}
a = \frac{1}{2} && b= \frac{1}{2} && c = \frac{1}{4} \\
\end{align}
```

$$a = \frac{1}{2} \qquad b = \frac{1}{2} \qquad c = \frac{1}{4}$$

In [9]:
```
%%latex
$e^{i\pi} + 1 = 0
$
```

$$e^{i\pi} + 1 = 0$$

Notebook MUST include…

‘THIS IS HOW’ = Code

‘THIS IS WHY’ = Comments

The Swivel Chair – Language Agnostic Work

In [11]:
```
logit femp mune und5
```

```
Iteration 0:   log likelihood = -1027.2309
Iteration 1:   log likelihood = -879.88806
Iteration 2:   log likelihood = -878.68101
Iteration 3:   log likelihood = -878.67998
Iteration 4:   log likelihood = -878.67998

Logistic regression                             Number of obs     =      1,580
                                                LR chi2(2)        =     297.10
                                                Prob > chi2       =     0.0000
Log likelihood = -878.67998                     Pseudo R2         =     0.1446

------------------------------------------------------------------------------
        femp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mune |  -1.703308   .2358489    -7.22   0.000    -2.165563   -1.241053
        und5 |  -1.733521   .1221909   -14.19   0.000    -1.973011   -1.494031
       _cons |   1.306829   .0744154    17.56   0.000     1.160978    1.452681
------------------------------------------------------------------------------
```

In [3]:
```
mylogit <- glm(femp ~ mune + und5, data = mydata, family = "binomial")

summary(mylogit)
```

```
Call:
glm(formula = femp ~ mune + und5, family = "binomial", data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7586  -1.0024   0.6922   0.6922   2.1177

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.30683    0.07442  17.561  < 2e-16 ***
mune        -1.70331    0.23585  -7.222 5.12e-13 ***
und5        -1.73352    0.12219 -14.187  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2054.5  on 1579  degrees of freedom
Residual deviance: 1757.4  on 1577  degrees of freedom
AIC: 1763.4
```

In [6]:
```
independentVar = ['mune', 'und5', 'Int']
logReg = sm.Logit(df['femp'] , df[independentVar])
answer = logReg.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.556127
         Iterations 5
```

the results are in the oject "answer"

In [9]:
```
answer.summary()
```

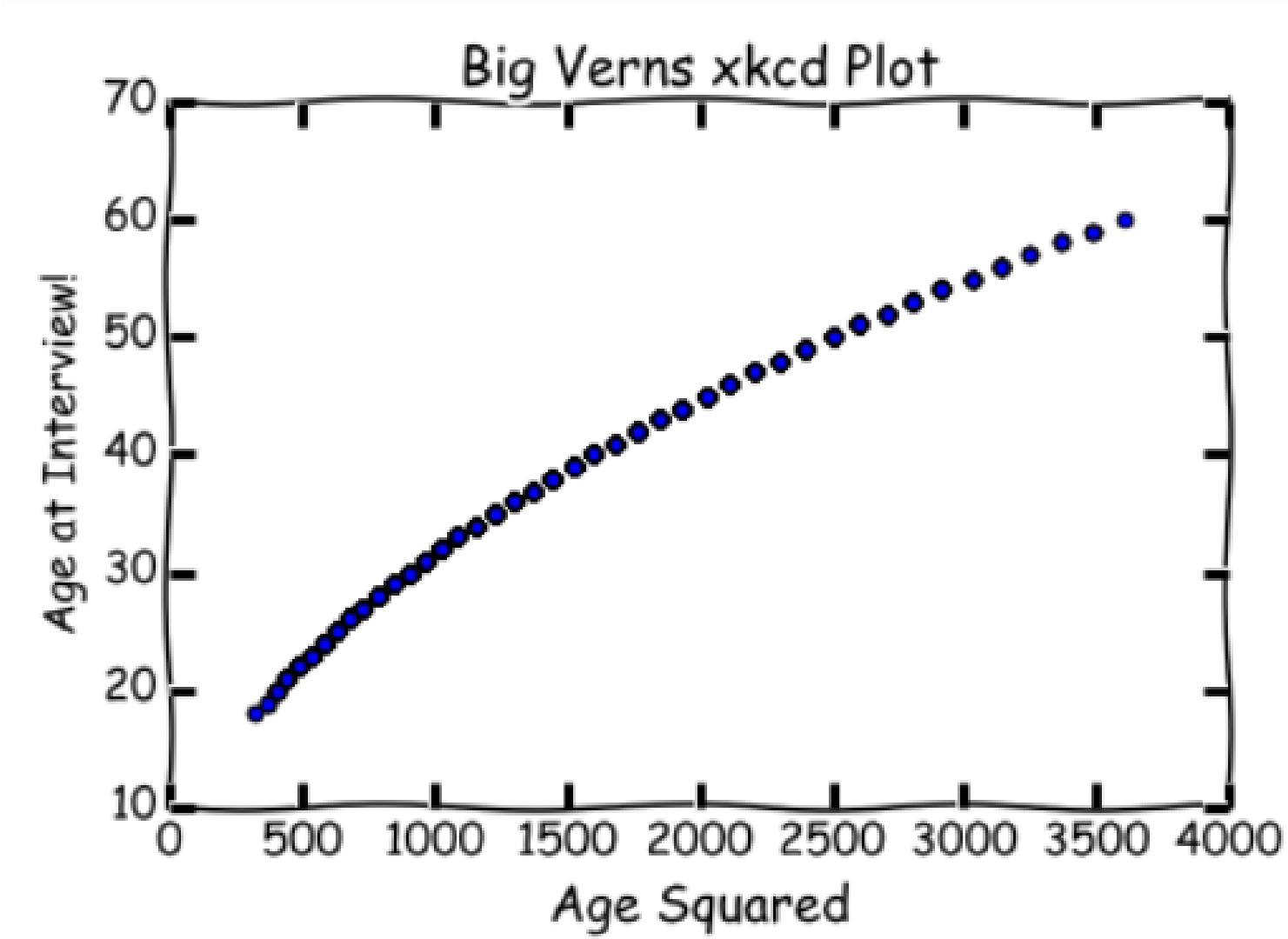
Out[9]:

Logit Regression Results

Dep. Variable:	femp	**No. Observations:**	1580
Model:	Logit	**Df Residuals:**	1577
Method:	MLE	**Df Model:**	2
Date:	Fri, 14 Oct 2016	**Pseudo R-squ.:**	0.1446
Time:	10:13:23	**Log-Likelihood:**	-878.68
converged:	True	**LL-Null:**	-1027.2
		LLR p-value:	3.056e-65

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
mune	-1.7033	0.236	-7.222	0.000	-2.166 -1.241
und5	-1.7335	0.122	-14.187	0.000	-1.973 -1.494
Int	1.3068	0.074	17.561	0.000	1.161 1.453

Rich Visual Outputs

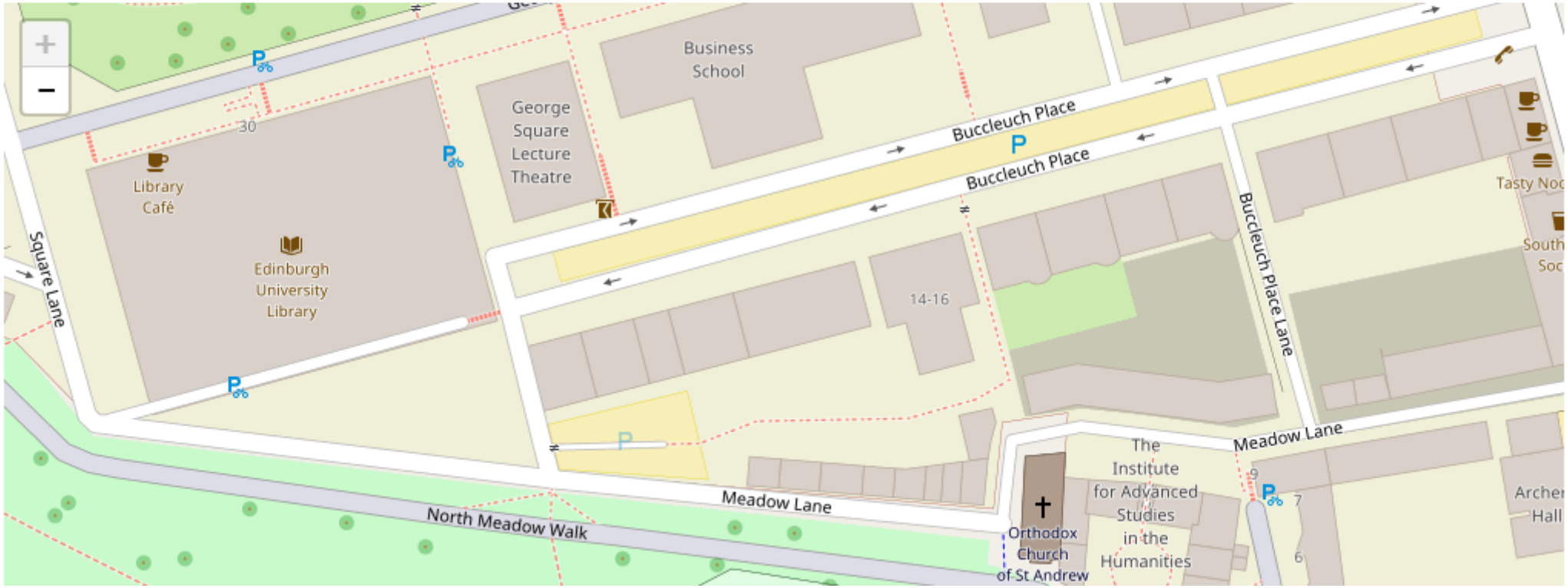


Another inventive use of the wemp dataset

Using an open street map

I've recently moved to a more commodious office in Buccleuch Place. Here is an example of an open source map on my new *hood*.

In [5]:
```
from ipyleaflet import Map
Map(center=[55.942535, -3.187269], zoom=20)
```



Some Points of Caution

- Easy to install but dependencies can be complex
- Windows 10, university systems etc. conspire against
- Open source = less help
- Stack Overflow, blogs etc. assume low-level programming skills

A Comprehensive Research Object

The British Journal of Sociology 2019 Volume 70 Issue 1

BJS

An investigation of social class inequalities in general cognitive ability in two British birth cohorts[1]

Roxanne Connelly and Vernon Gayle

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Filename	Description
bjos12343-sup-0001-suppinfo1.ipynb 821.7 KB	**Appendix S1:** Jupyter Notebook file
bjos12343-sup-0002-suppinfo2.pdf 9.3 MB	**Appendix S2:** Jupyter Notebook file in PDF
bjos12343-sup-0003-suppinfo3.docx 38.7 KB	**Online Supplement:** Additional Analyses

Please note: The publisher is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing content) should be directed to the corresponding author for the article.

Connelly, R. and Gayle, V., 2019. An investigation of social class inequalities in general cognitive ability in two British birth cohorts. *The British journal of sociology*, *70*(1), pp.90-108.

RoxanneConnelly / Social-Class-Inequalities-in-General-Cognitive-Ability-in-Two-British-Birth-Cohorts

No description, website, or topics provided.

README.md

An investigation of Social Class Inequalities in General Cognitive Ability in Two British Birth Cohorts

British Journal of Sociology

Roxanne Connelly (R.Connelly@warwick.ac.uk)

Vernon Gayle (vernon.gayle@ed.ac.uk)

This repository hosts a Jupyter Notebook which accompanies the paper above. Details of how to access the required data are provided in the notebook.

https://tinyurl.com/qnl38m7

https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#social-data

Sociology: An investigation of Social Class Inequalities in General Cognitive Ability in Two British Birth Cohorts. Preprint in SocArXiv, December 2017. doi: 10.17605/OSF.IO/SZXDM. Notebook and materials at: OSF, GitHub, nbviewer

Jupyter Notebooks

Office of Planning & Analysis
UC Berkeley
March 2020

Professor Vernon Gayle

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

NCRM National Centre for Research Methods