Jupyter Notebooks a Quick-Step Towards Literate Computing and Reproducible Research

Q – Step Edinburgh, January 2017

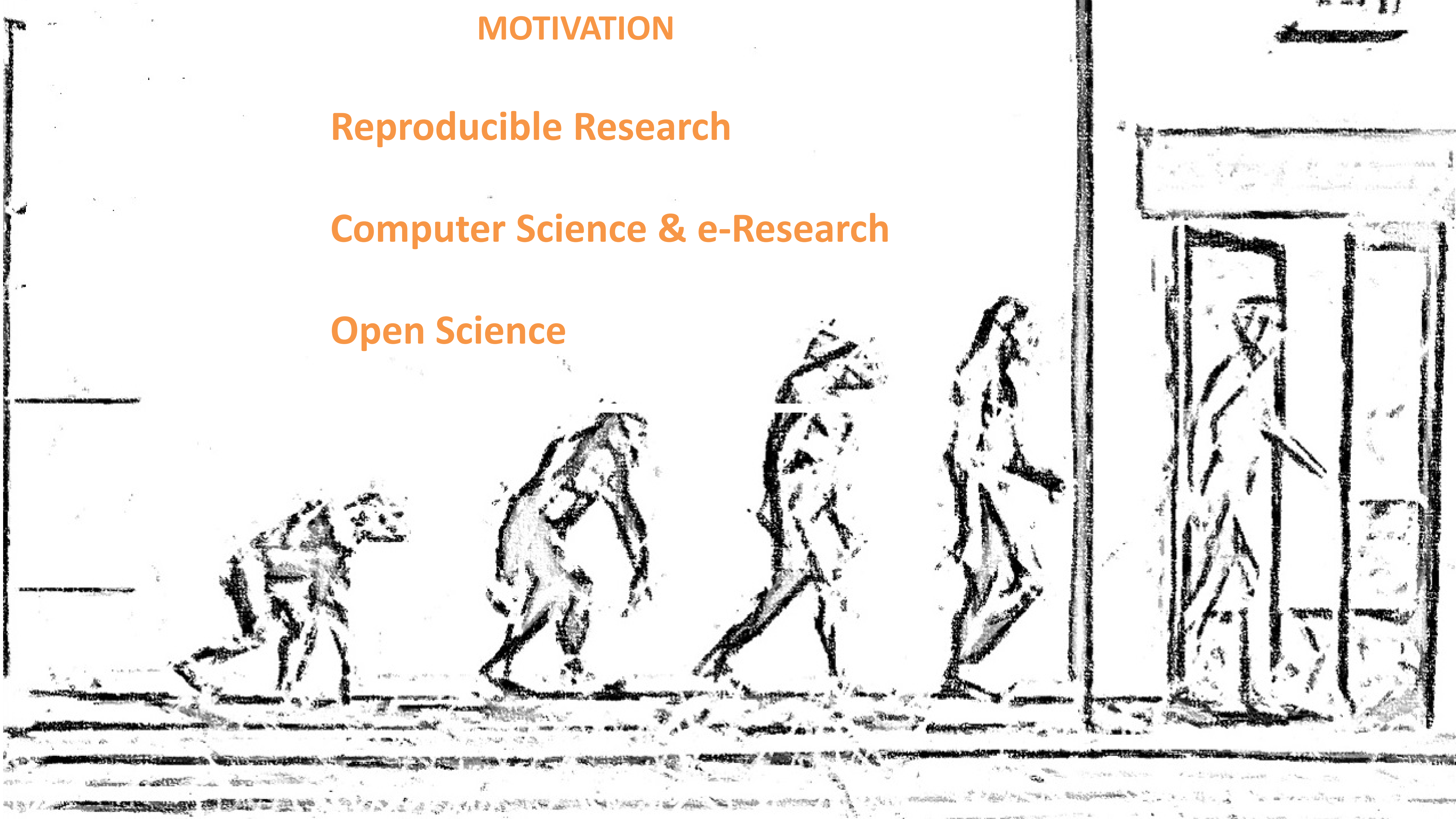
Vernon Gayle
University of Edinburgh
@profbigvern

MOTIVATION

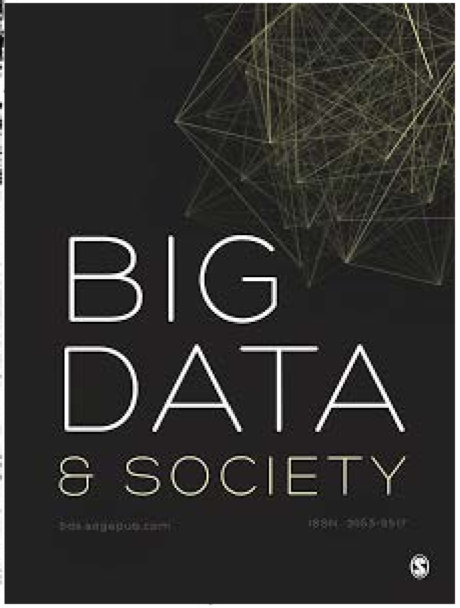
Reproducible Research

Computer Science & e-Research

Open Science



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).

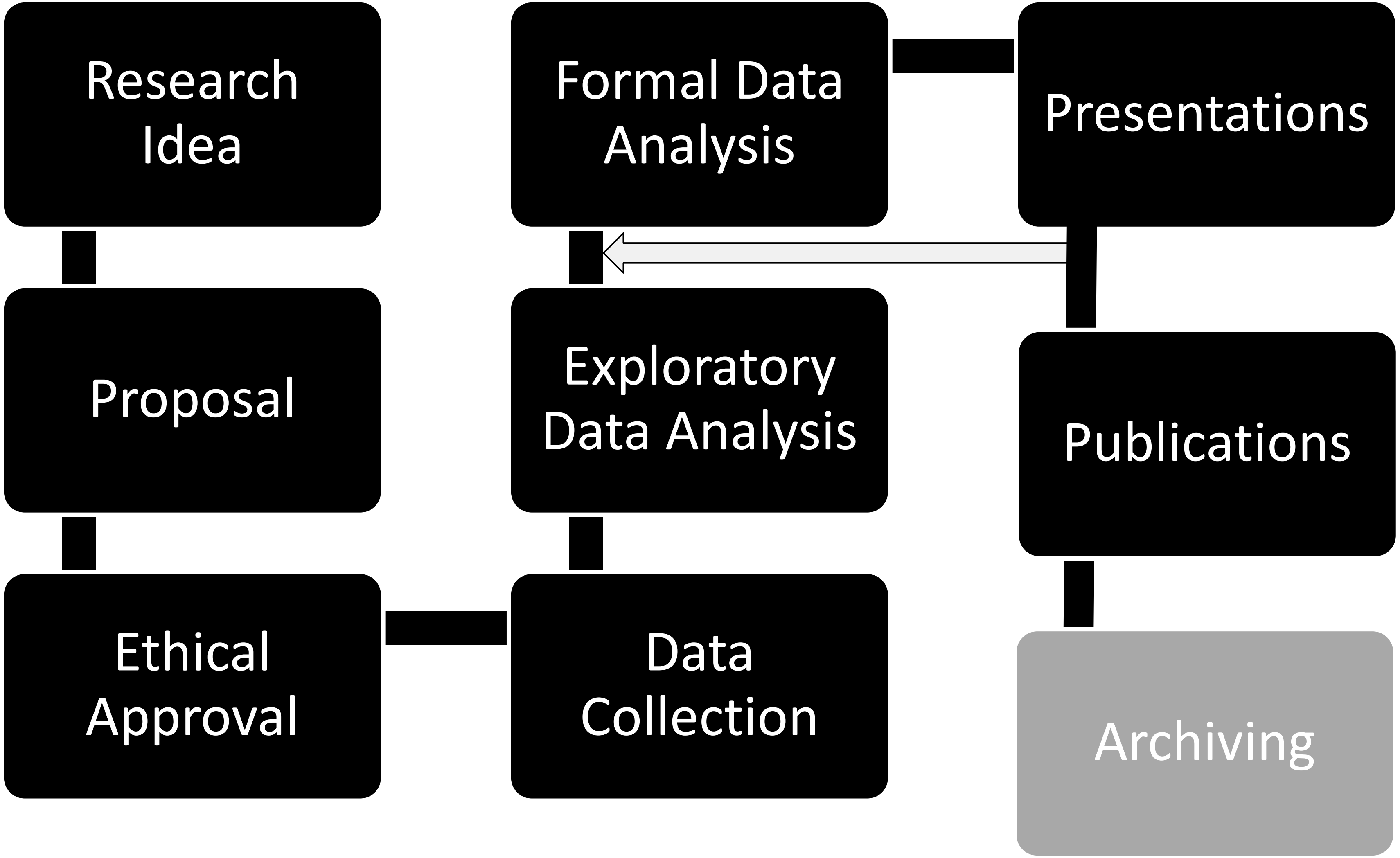


Structure of this Session

- Workflow
- Literate computing
- Notebooks in research
- Jupyter in general
- Jupyter demonstration
- Concluding remarks

A tigger warning – there will be live software demonstrated and things might get bouncy and could potentially go wrong!

The Workflow



Research Idea

Proposal

Ethical Approval

Data Collection

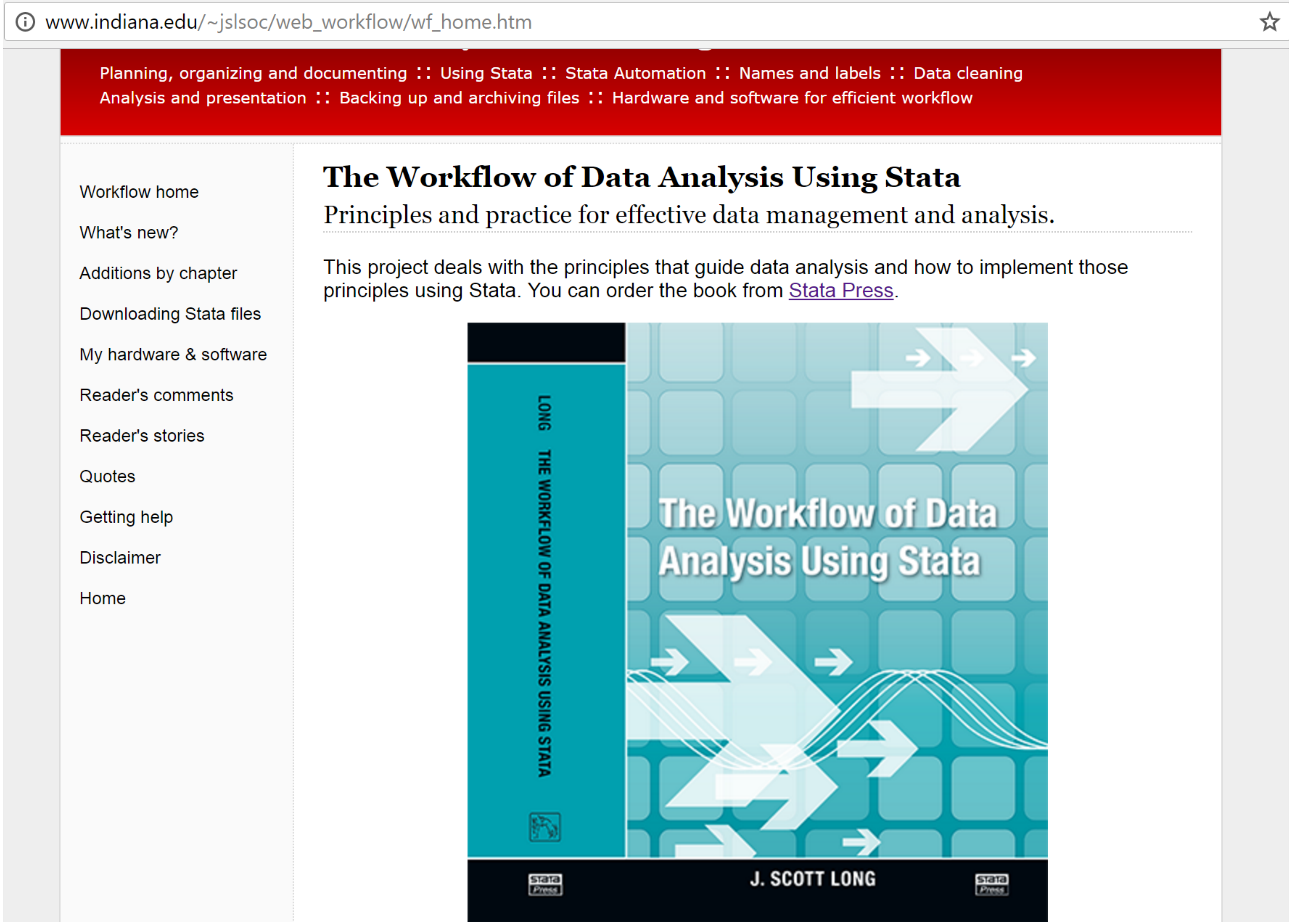
Exploratory Data Analysis

Formal Data Analysis

Presentations

Publications

Archiving



www.indiana.edu/~jslsoc/web_workflow/wf_home.htm

Planning, organizing and documenting :: Using Stata :: Stata Automation :: Names and labels :: Data cleaning
Analysis and presentation :: Backing up and archiving files :: Hardware and software for efficient workflow

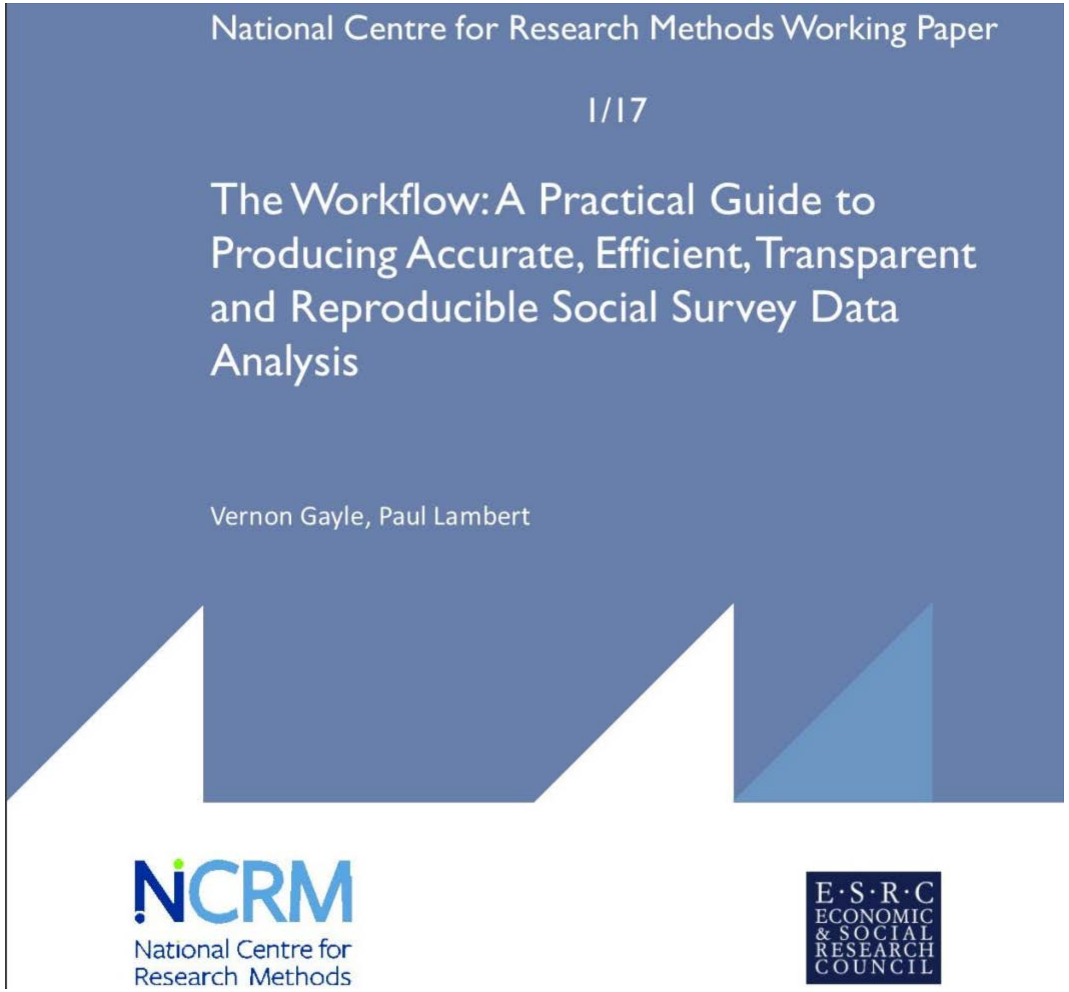
- Workflow home
- What's new?
- Additions by chapter
- Downloading Stata files
- My hardware & software
- Reader's comments
- Reader's stories
- Quotes
- Getting help
- Disclaimer
- Home

The Workflow of Data Analysis Using Stata

Principles and practice for effective data management and analysis.

This project deals with the principles that guide data analysis and how to implement those principles using Stata. You can order the book from Stata Press.

LONG THE WORKFLOW OF DATA ANALYSIS USING STATA
The Workflow of Data Analysis Using Stata
J. SCOTT LONG
Stata Press



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

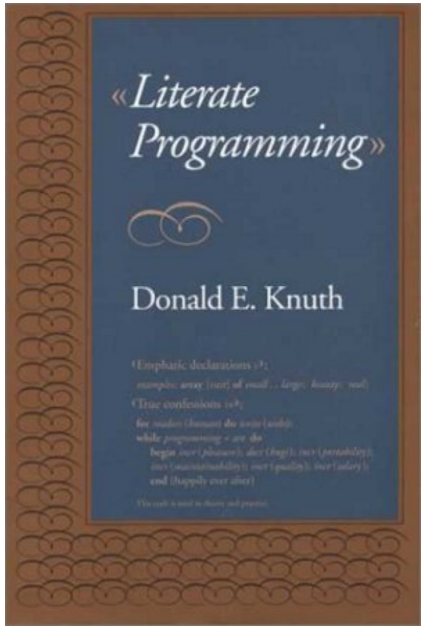
http://eprints.ncrm.ac.uk/4000/

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/

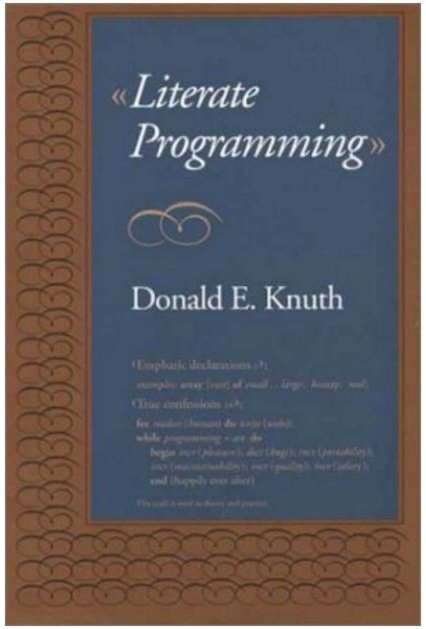


Knuth says

Treat your program as literature

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

http://tinyurl.com/he5aagf



Knuth says

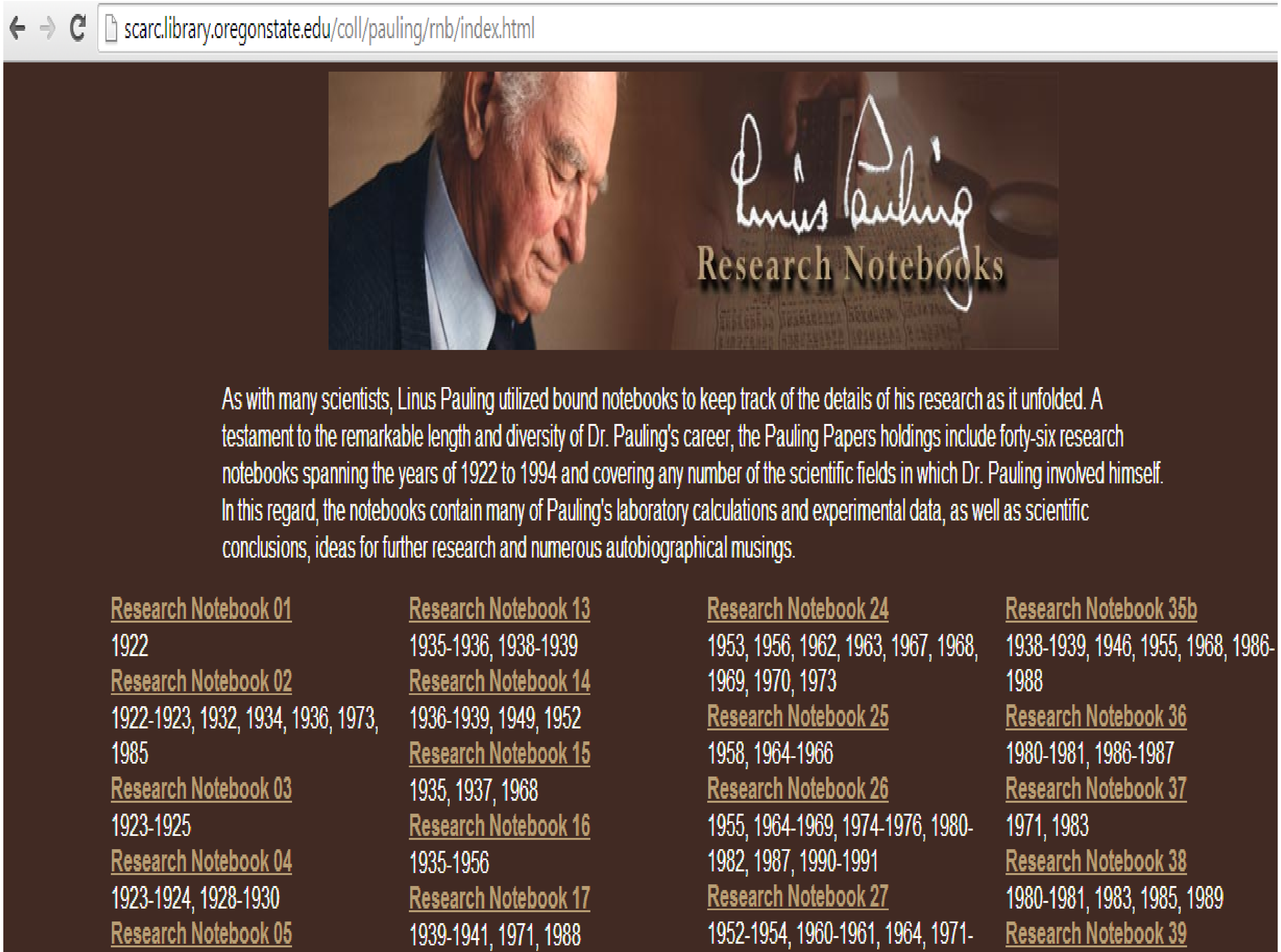
Treat your program as literature

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 may I will understand my own program better

http://tinyurl.com/he5aagf



Linus Pauling

Research Notebooks

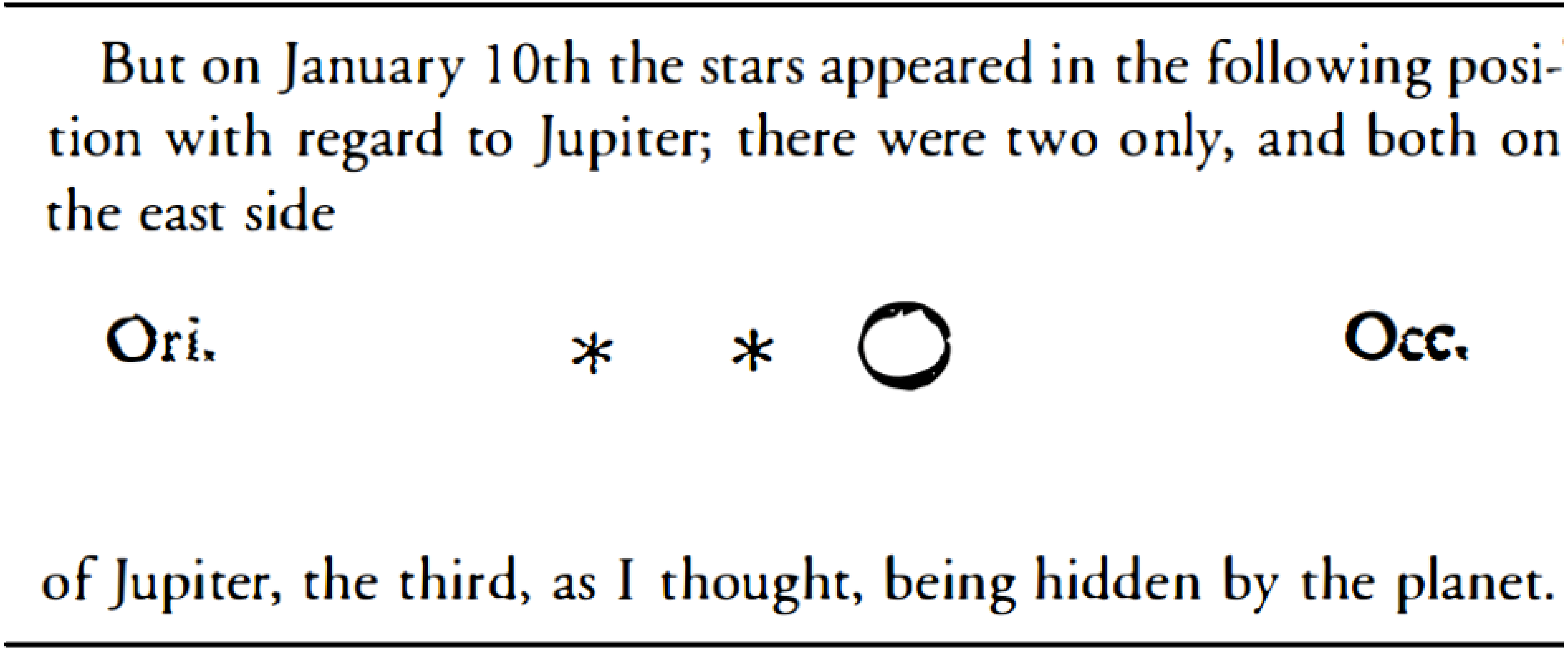
As with many scientists, Linus Pauling utilized bound notebooks to keep track of the details of his research as it unfolded. A testament to the remarkable length and diversity of Dr. Pauling's career, the Pauling Papers holdings include forty-six research notebooks spanning the years of 1922 to 1994 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

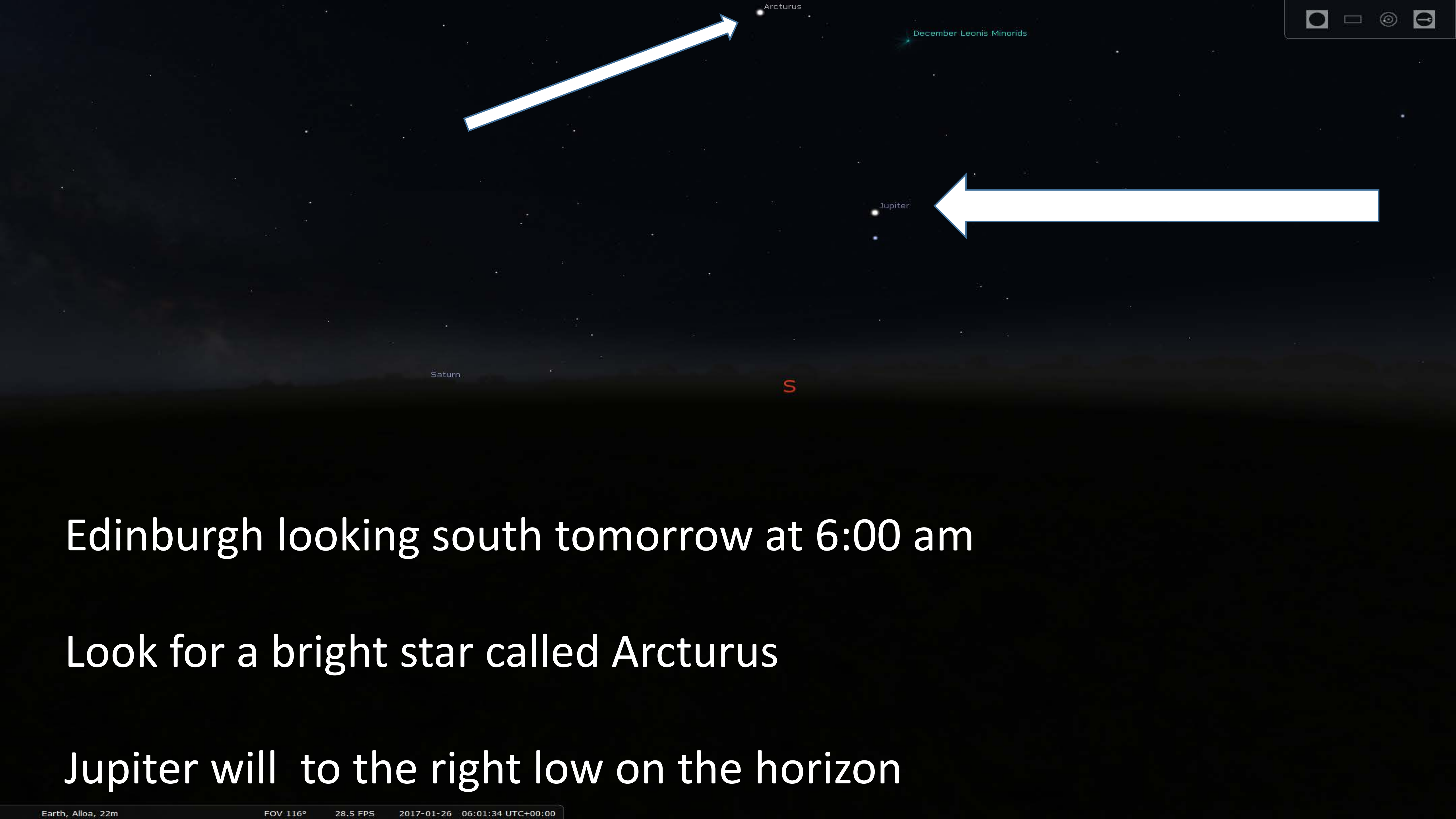


But on January 10th the stars appeared in the following position with regard to Jupiter; there were two only, and both on the east side

Ori. * * ○ Occ.

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

Sidereus Nuncius (Galilei 1610)



Earth, Alloa, 22m FOV 116° 28.5 FPS 2017-01-26 06:01:34 UTC+00:00

Edinburgh looking south tomorrow at 6:00 am

Look for a bright star called Arcturus

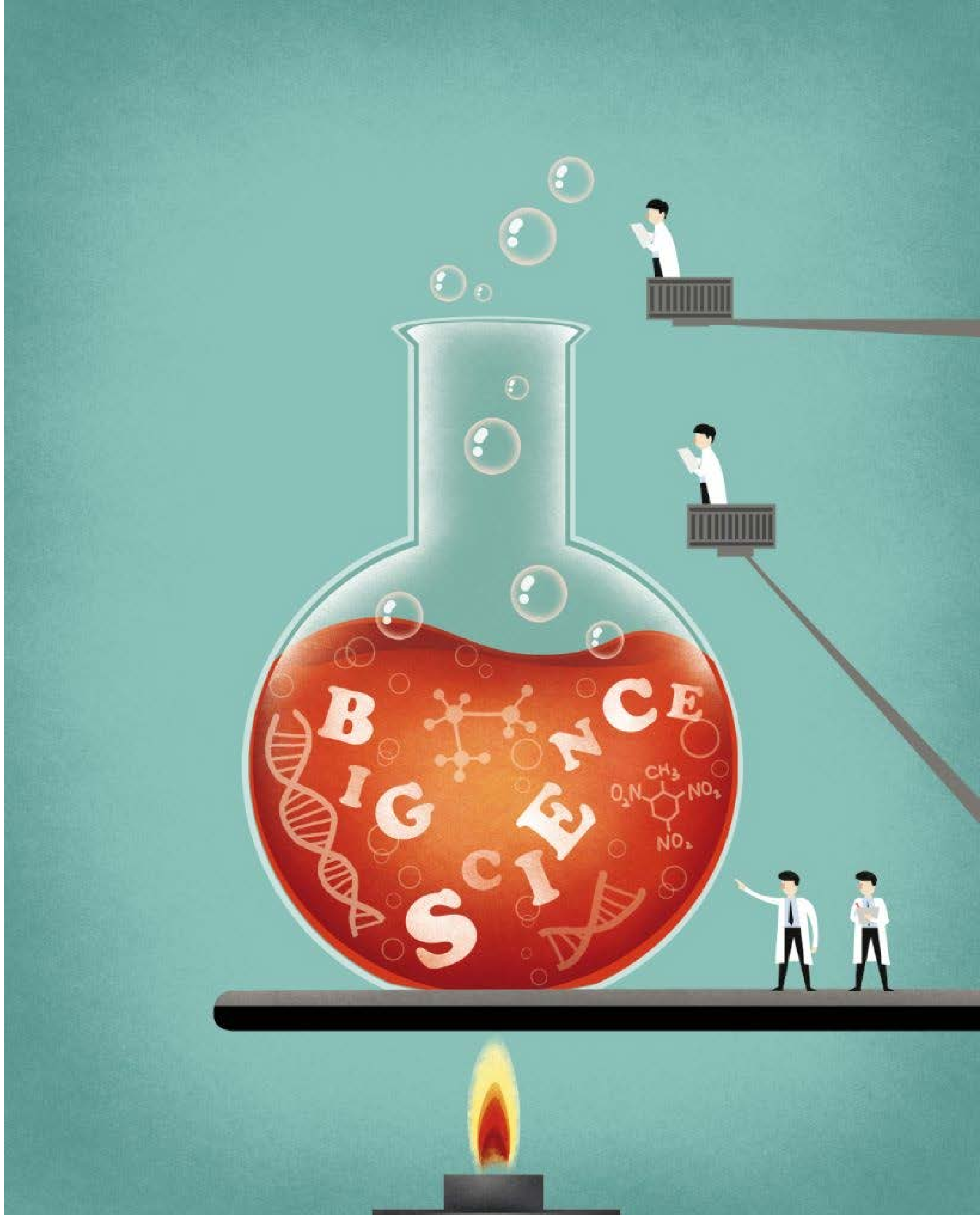
Jupiter will to the right low on the horizon

Juila, Python and R almost spell JuPyteR

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

https://jupyter.org/

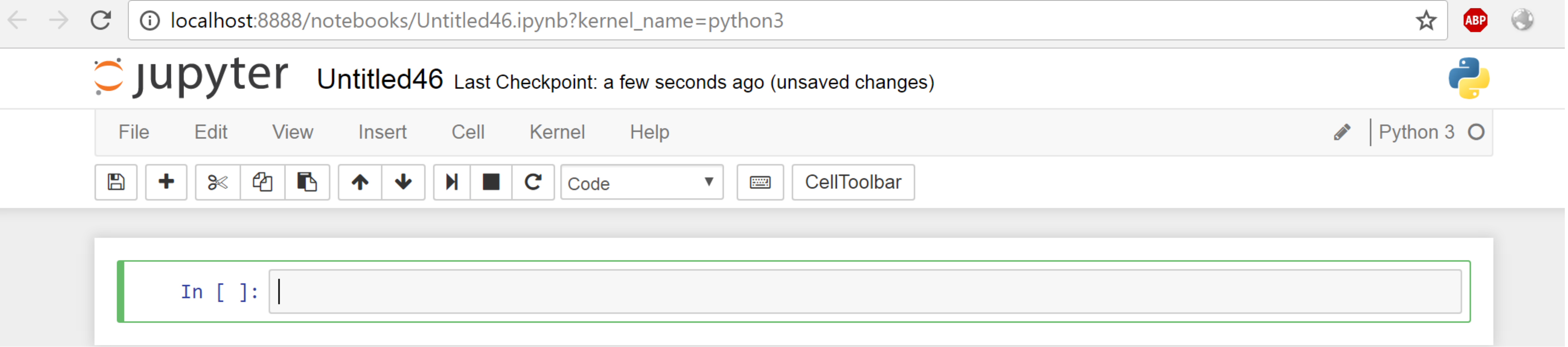
jupyter


BIG SCIENCE
CH3
O2N
NO2
NO2

https://www.youtube.com/watch?v=BmHPoBpZoJ4

- Easy documentation alongside research code

- 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

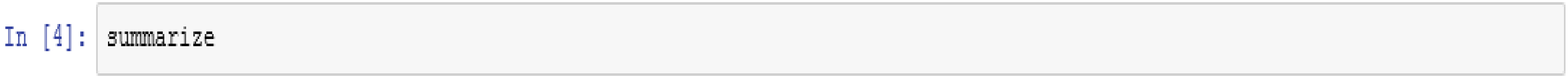

localhost:8888/notebooks/Untitled46.ipynb?kernel_name=python3
jupyter
Untitled46
Last Checkpoint: a few seconds ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Help
Python 3
Code
CellToolbar
In [ ]:

Open source web application

Creates documents which include live code, output and explanatory text

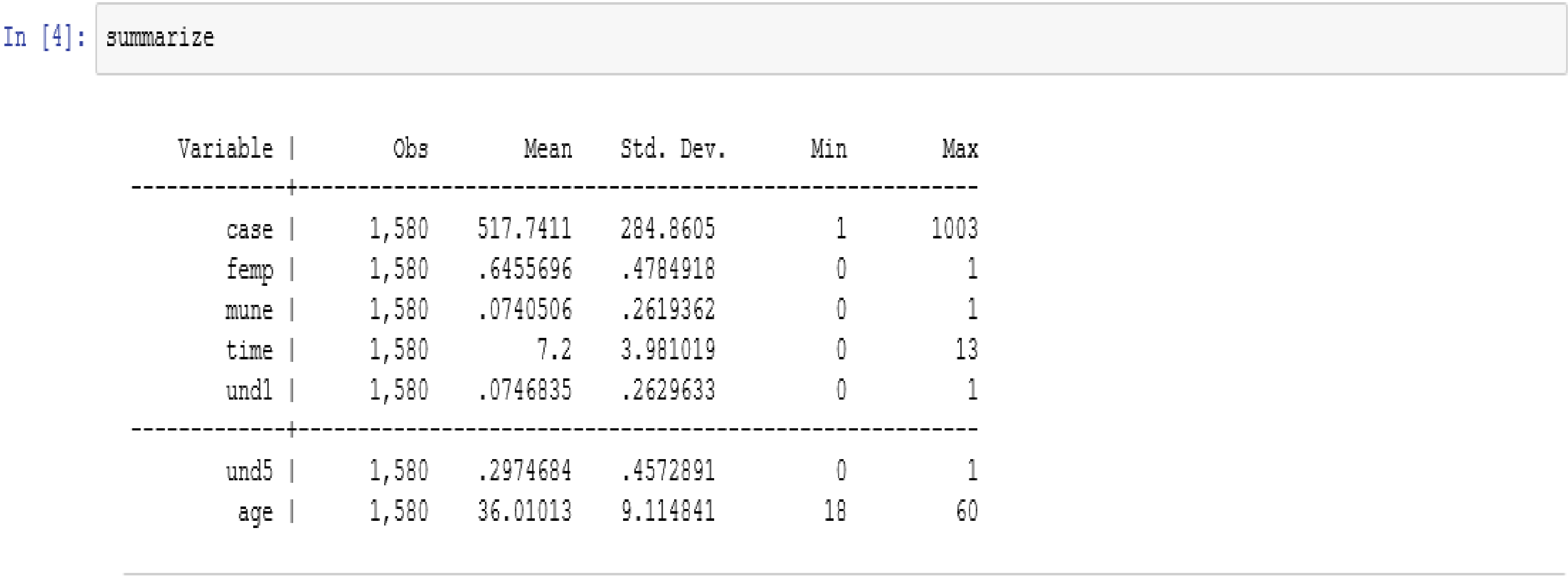
Single platform for the complete workflow

Code



```
summarize
```

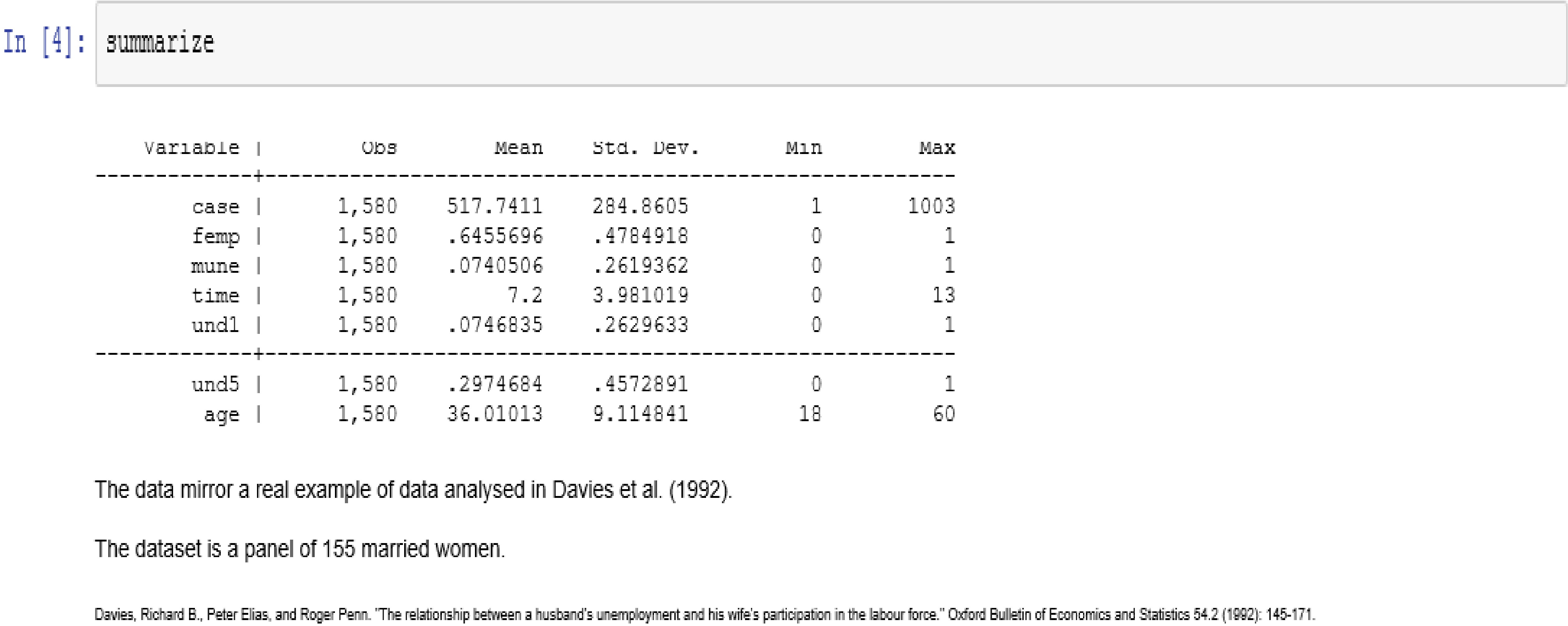
Output



```
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

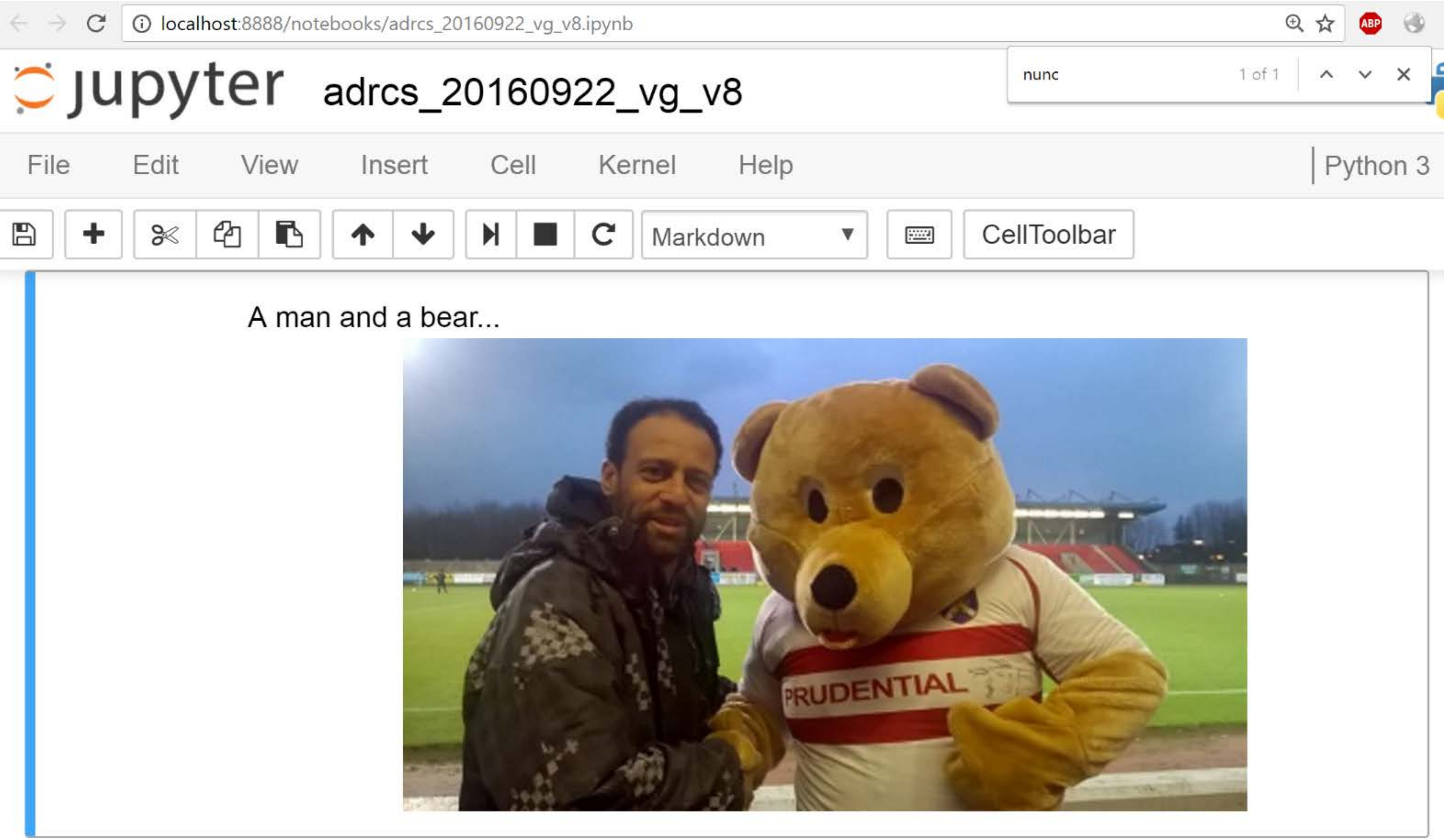
- It’s quickly becoming the writing standard in some academic areas and in science

- 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)

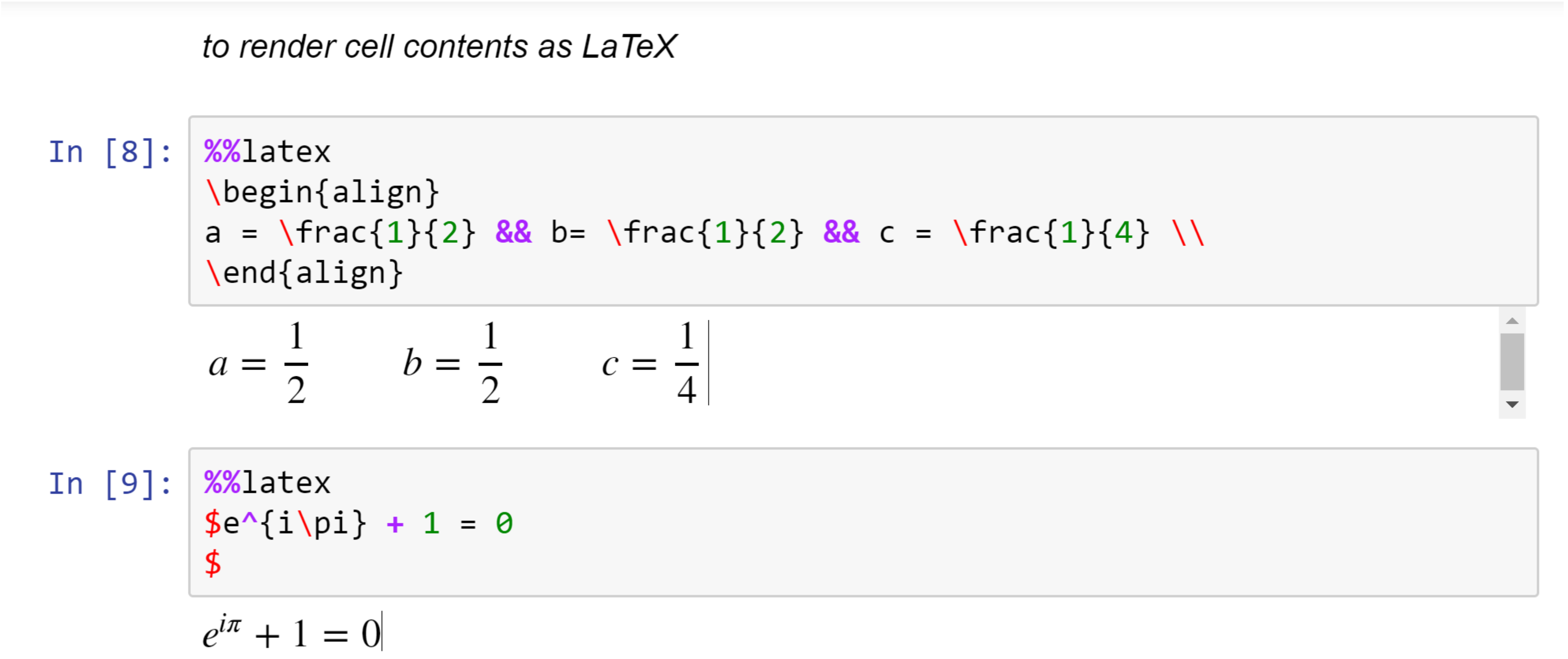
Images within the notebook cell...



A man and a bear...

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}
```

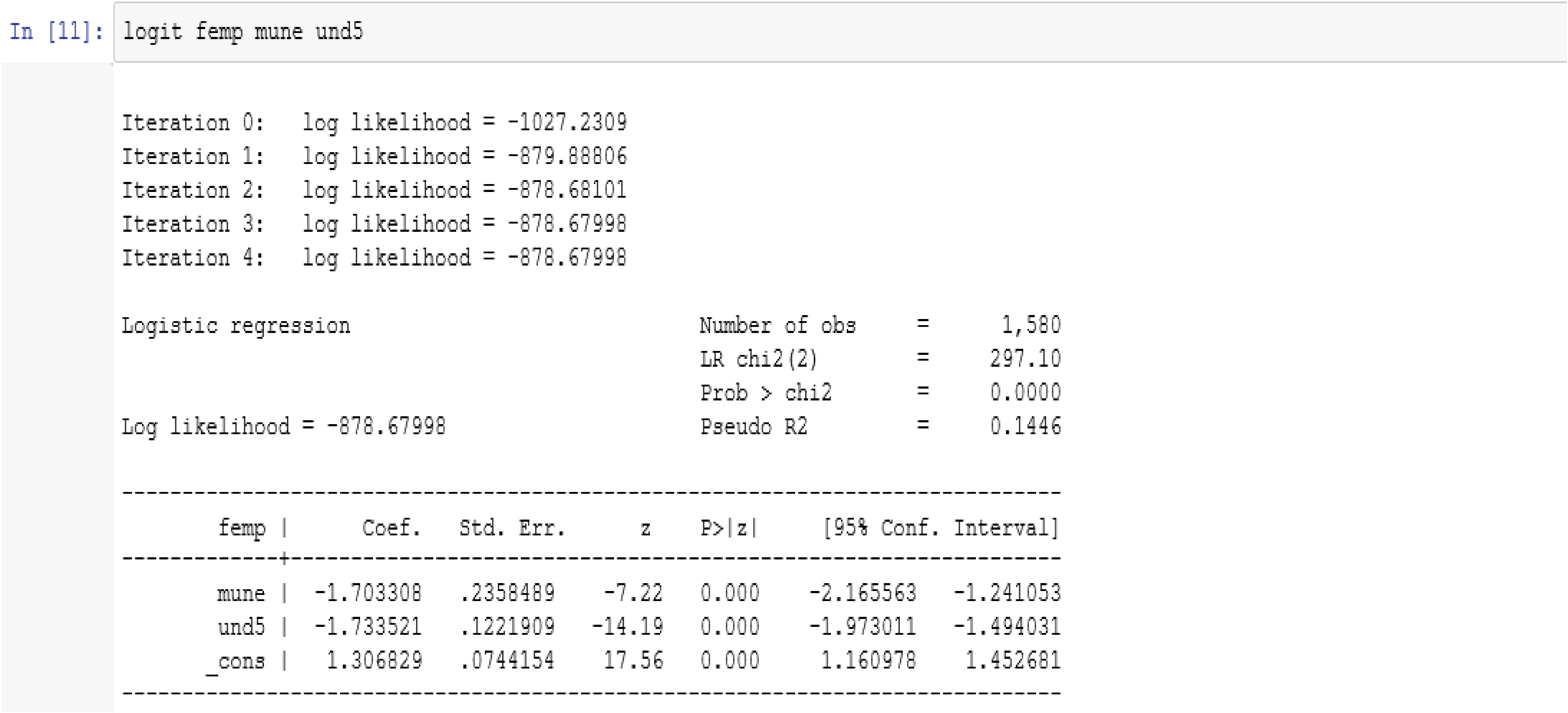
$$a = \frac{1}{2} \qquad b = \frac{1}{2} \qquad c = \frac{1}{4}$$

In [9]:

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

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

The Swivel Chair – Language Agnostic Work

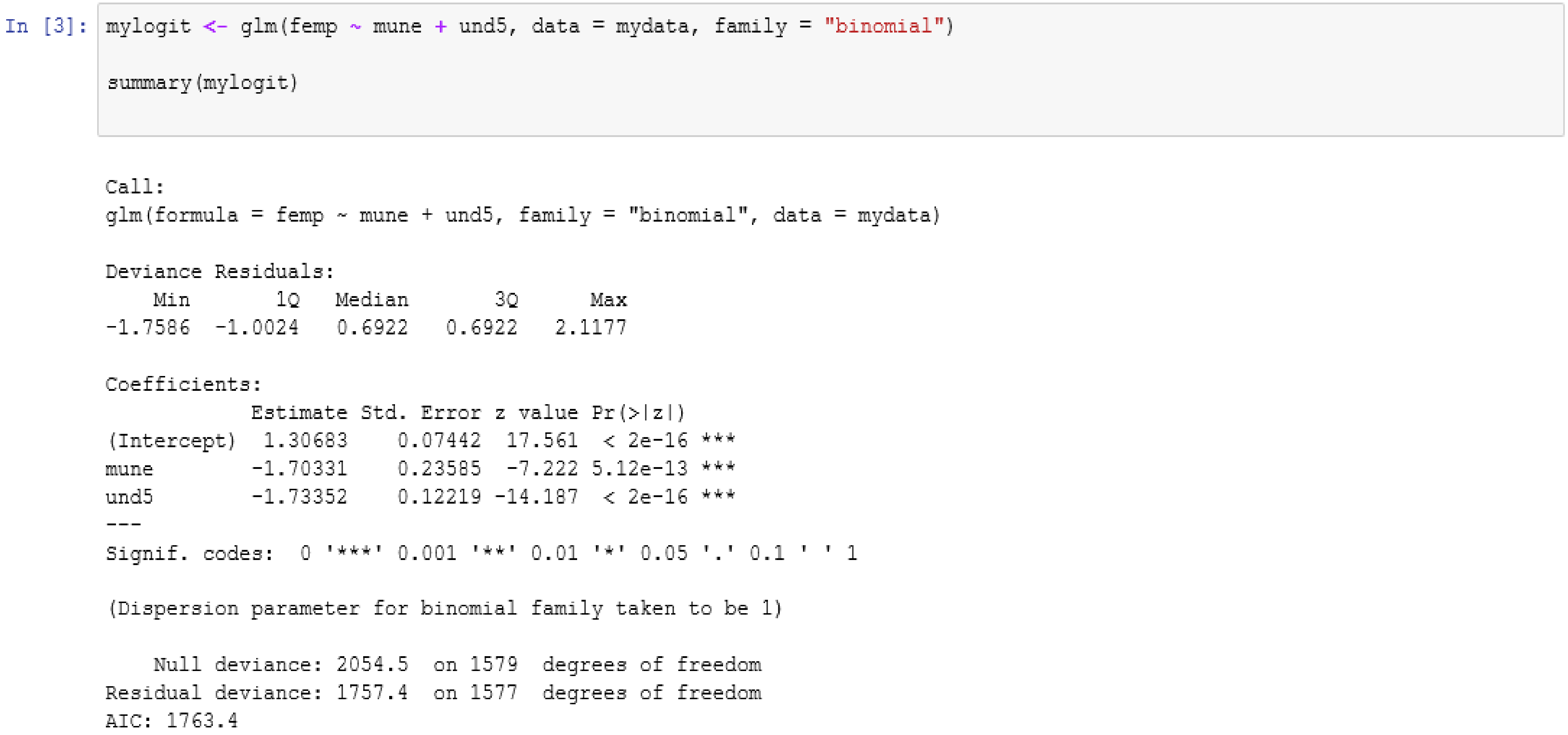


```
In [11]: logit femp mune und5
```

```
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)
```

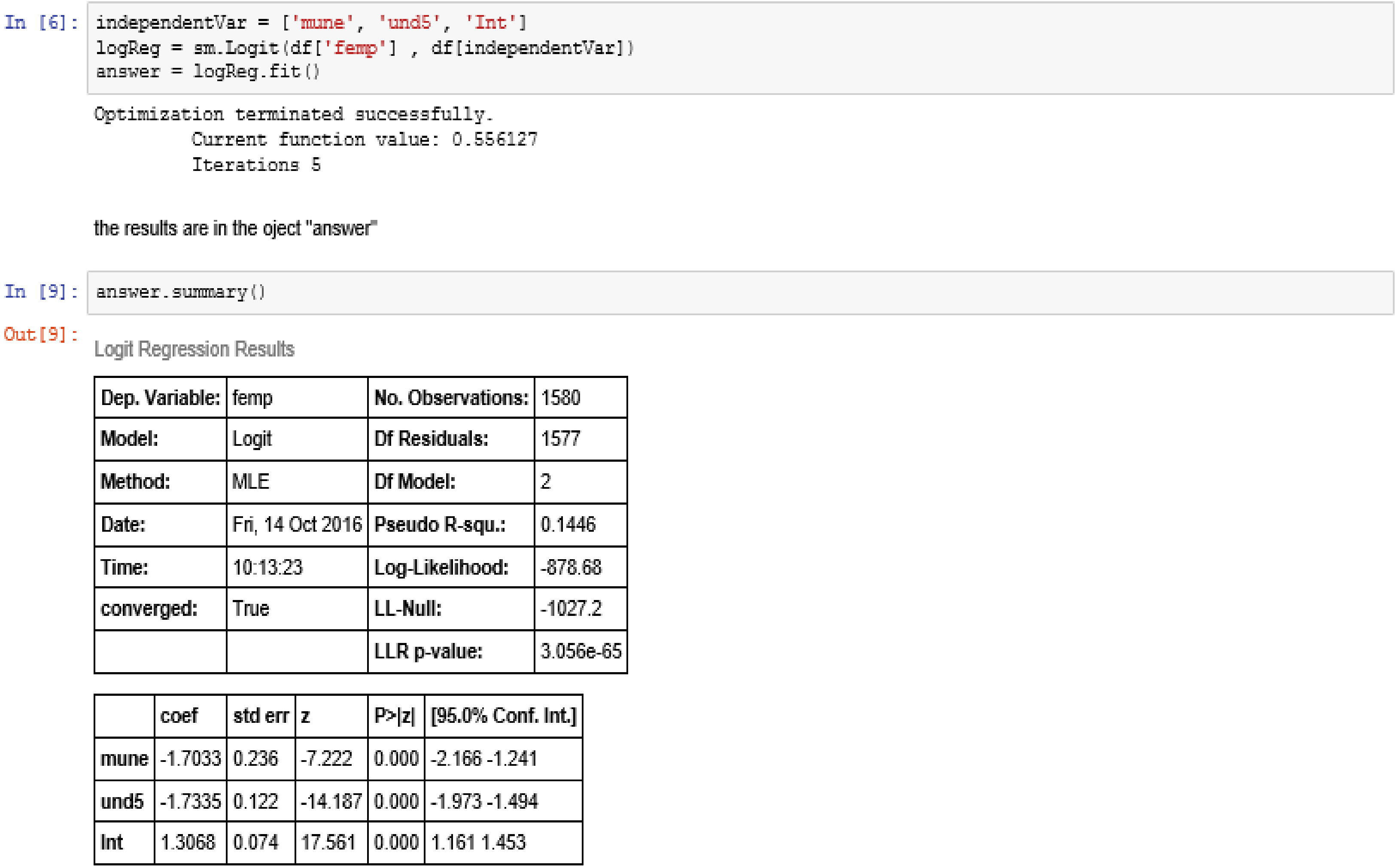
```
Call:
glm(formula = femp ~ mune + und5, family = "binomial", data = mydata)

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()
```

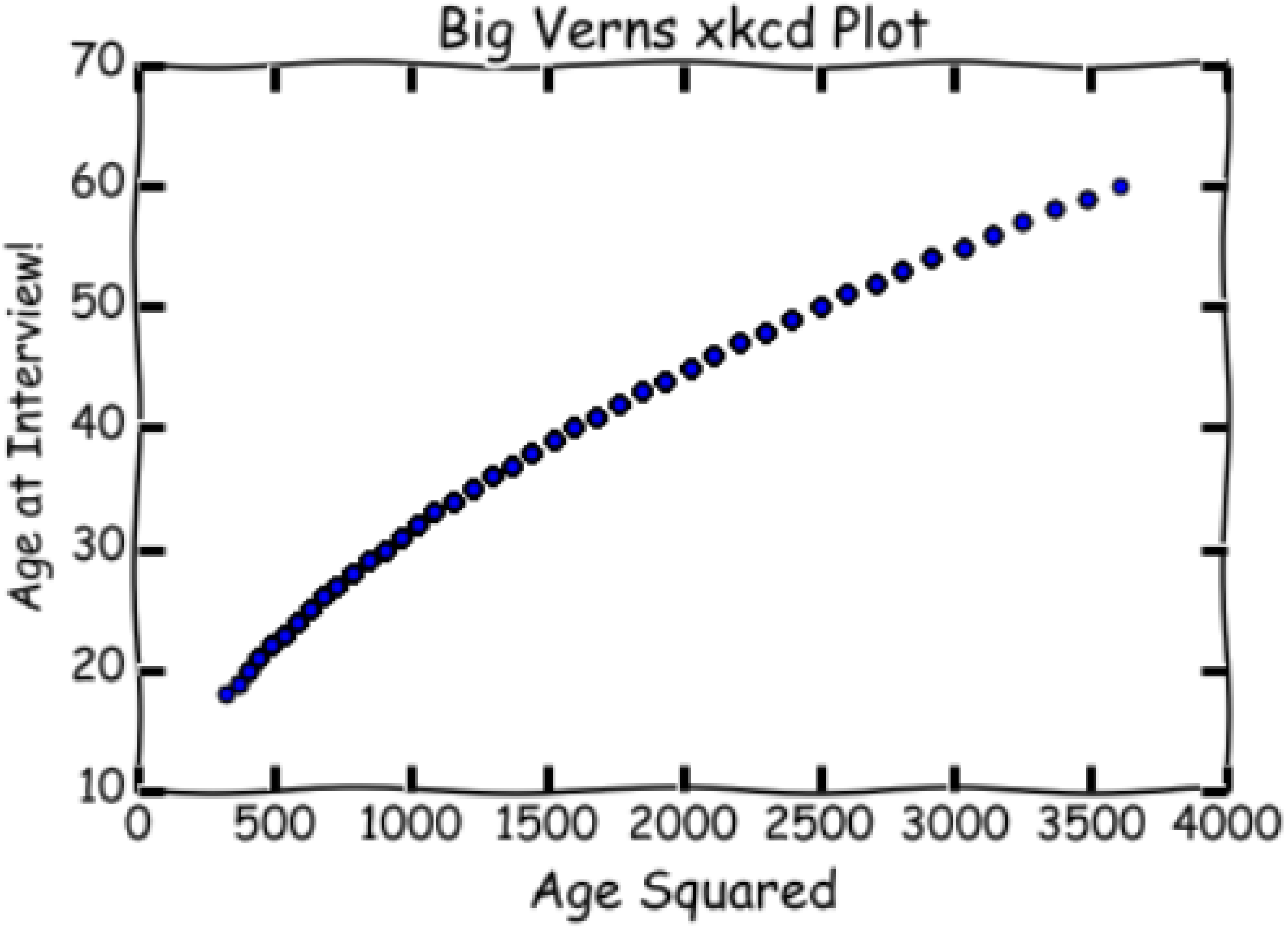
Out[9]:

Logit Regression Results

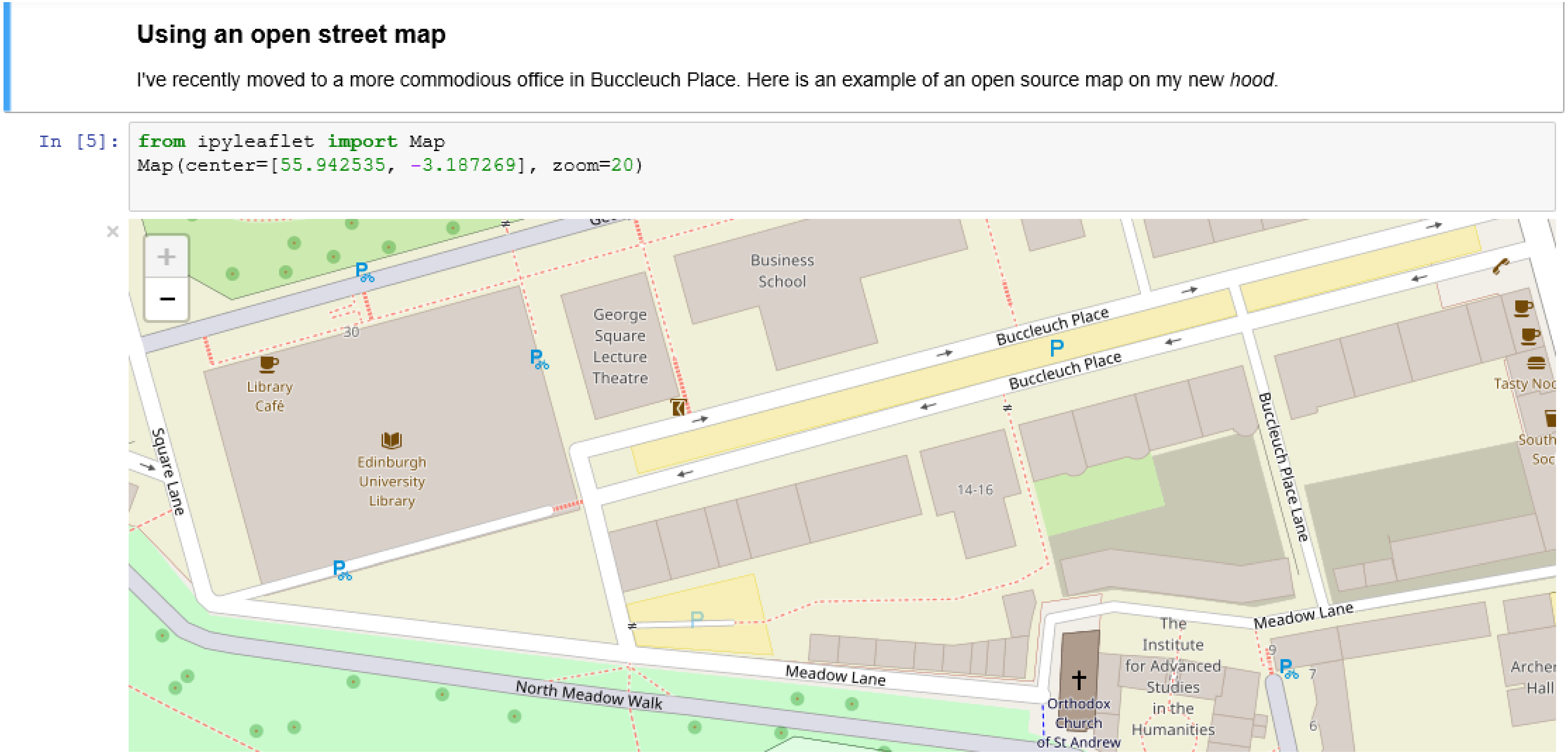
Dep. Variable:	femp	**No. Observations:**	1580
Model:	Logit	**Df Residuals:**	1577
Method:	MLE	**Df Model:**	2
Date:	Fri, 14 Oct 2016	**Pseudo R-squ.:**	0.1446
Time:	10:13:23	**Log-Likelihood:**	-878.68
converged:	True	**LL-Null:**	-1027.2
		LLR p-value:	3.056e-65

	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



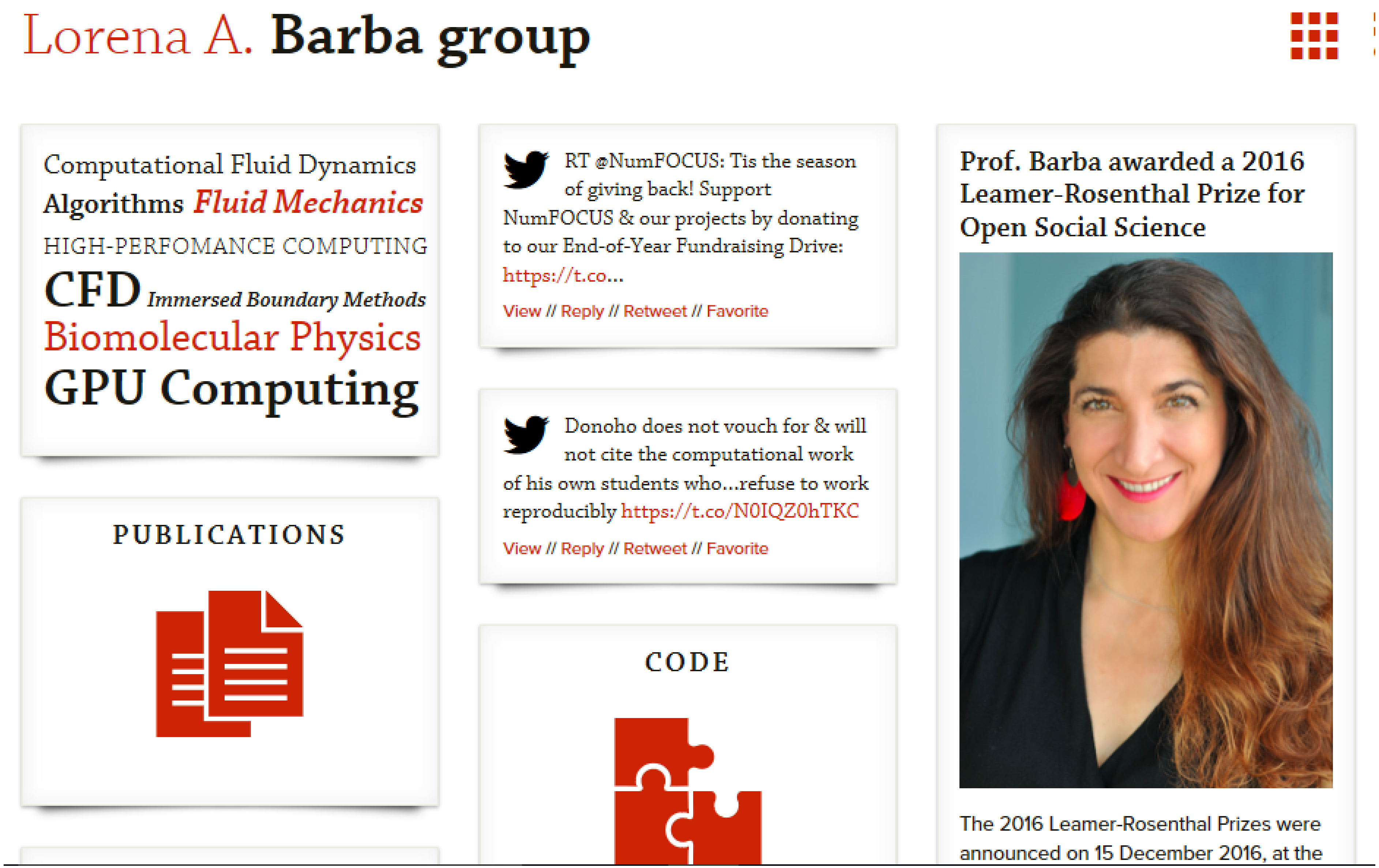
Another inventive use of the wemp dataset



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)
```



Lorena A. Barba group

Computational Fluid Dynamics
Algorithms *Fluid Mechanics*
HIGH-PERFOMANCE COMPUTING
CFD *Immersed Boundary Methods*
Biomolecular Physics
GPU Computing

PUBLICATIONS

RT @NumFOCUS: Tis the season of giving back! Support NumFOCUS & our projects by donating to our End-of-Year Fundraising Drive: https://t.co...

View // Reply // Retweet // Favorite

Donoho does not vouch for & will not cite the computational work of his own students who...refuse to work reproducibly https://t.co/N0IQZ0hTKC

View // Reply // Retweet // Favorite

CODE

Prof. Barba awarded a 2016 Leamer-Rosenthal Prize for Open Social Science

The 2016 Leamer-Rosenthal Prizes were announced on 15 December 2016, at the

http://lorenabarba.com/

nbgrader

nbgrader is a tool that facilitates creating and grading assignments in the Jupyter notebook

It allows instructors to easily create notebook-based assignments that include both coding exercises and written free-responses

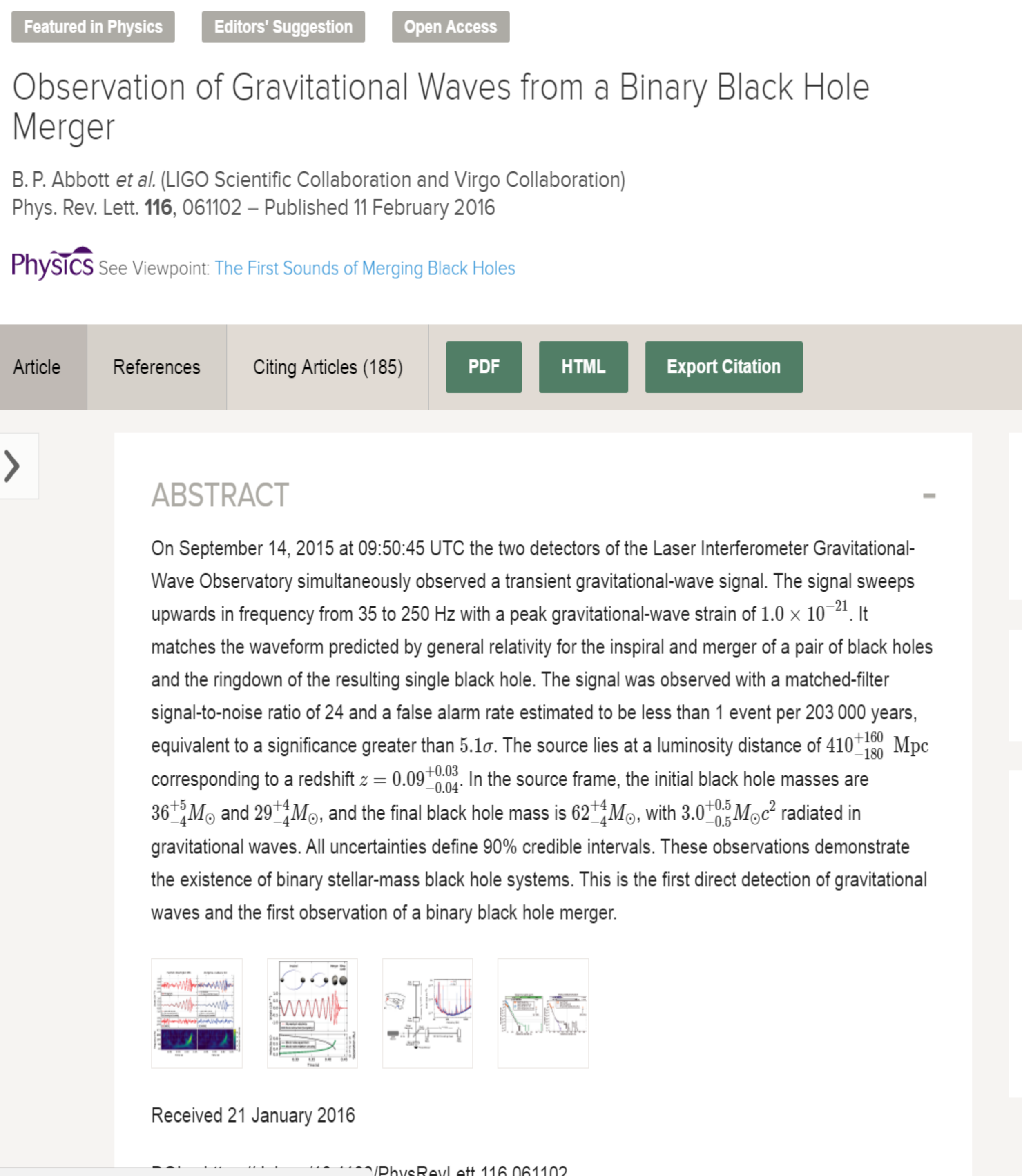
nbgrader then also provides a streamlined interface for quickly grading completed assignments

https://nbgrader.readthedocs.io/en/stable/

nbconvert

nbconvert converts notebooks to familiar formats e.g. PDF HTML LaTeX

Presentation, publishing, sharing and collaboration



Observation of Gravitational Waves from a Binary Black Hole Merger

B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration)

Phys. Rev. Lett. **116**, 061102 – Published 11 February 2016

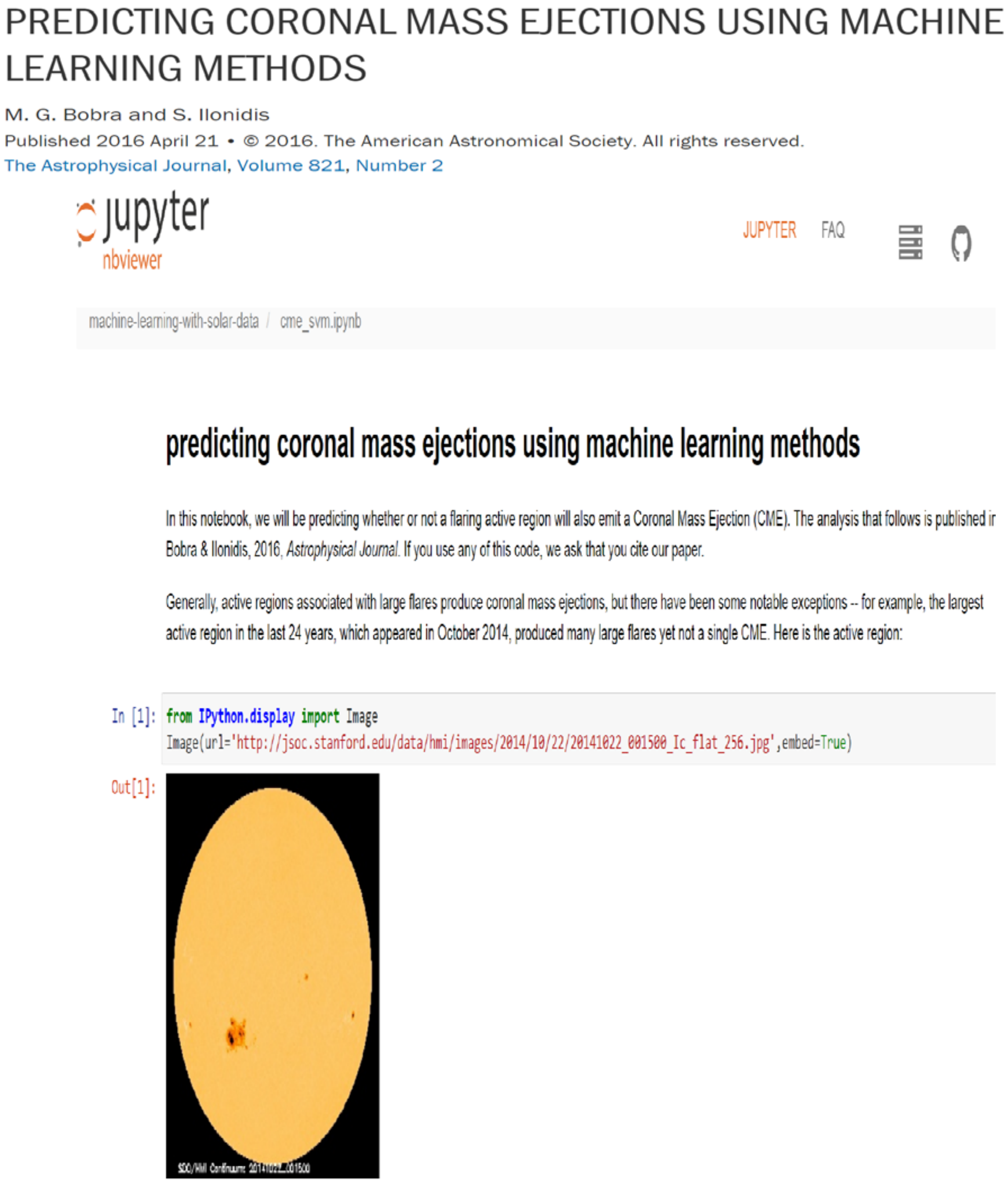
Physics See Viewpoint: The First Sounds of Merging Black Holes

ABSTRACT

On September 14, 2015 at 09:50:45 UTC the two detectors of the Laser Interferometer Gravitational-Wave Observatory simultaneously observed a transient gravitational-wave signal. The signal sweeps upwards in frequency from 35 to 250 Hz with a peak gravitational-wave strain of 1.0×10^{-21}. It matches the waveform predicted by general relativity for the inspiral and merger of a pair of black holes and the ringdown of the resulting single black hole. The signal was observed with a matched-filter signal-to-noise ratio of 24 and a false alarm rate estimated to be less than 1 event per 203 000 years, equivalent to a significance greater than 5.1σ. The source lies at a luminosity distance of $410^{+160}_{-180}\ \mathrm{Mpc}$ corresponding to a redshift $z = 0.09^{+0.03}_{-0.04}$. In the source frame, the initial black hole masses are $36^{+5}_{-4} M_\odot$ and $29^{+4}_{-4} M_\odot$, and the final black hole mass is $62^{+4}_{-4} M_\odot$, with $3.0^{+0.5}_{-0.5} M_\odot c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary black hole merger.

Received 21 January 2016

DOI: https://doi.org/10.1103/PhysRevLett.116.061102



PREDICTING CORONAL MASS EJECTIONS USING MACHINE LEARNING METHODS

M. G. Bobra and S. Ilonidis

Published 2016 April 21 • © 2016. The American Astronomical Society. All rights reserved.
The Astrophysical Journal, Volume 821, Number 2

predicting coronal mass ejections using machine learning methods

In this notebook, we will be predicting whether or not a flaring active region will also emit a Coronal Mass Ejection (CME). The analysis that follows is published ir Bobra & Ilonidis, 2016, *Astrophysical Journal*. If you use any of this code, we ask that you cite our paper.

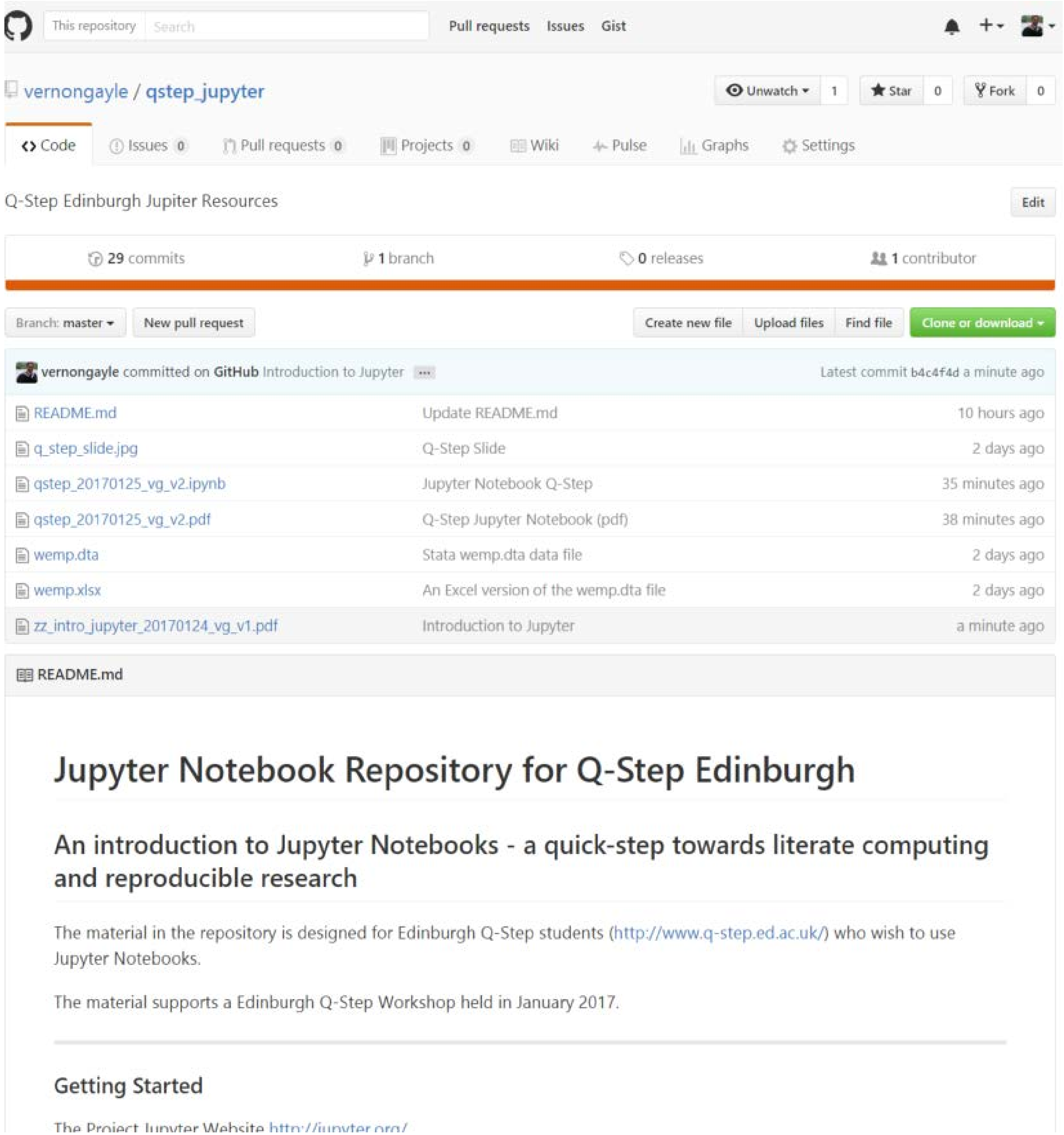
Generally, active regions associated with large flares produce coronal mass ejections, but there have been some notable exceptions -- for example, the largest active region in the last 24 years, which appeared in October 2014, produced many large flares yet not a single CME. Here is the active region:

```
In [1]: from IPython.display import Image
        Image(url='http://jsoc.stanford.edu/data/hmi/images/2014/10/22/20141022_001500_Ic_flat_256.jpg',embed=True)
Out[1]:
```

- 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

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



vernongayle / qstep_jupyter

Q-Step Edinburgh Jupiter Resources

29 commits | 1 branch | 0 releases | 1 contributor

vernongayle committed on GitHub Introduction to Jupyter ... Latest commit b4c4f4d a minute ago

README.md	Update README.md	10 hours ago
q_step_slide.jpg	Q-Step Slide	2 days ago
qstep_20170125_vg_v2.ipynb	Jupyter Notebook Q-Step	35 minutes ago
qstep_20170125_vg_v2.pdf	Q-Step Jupyter Notebook (pdf)	38 minutes ago
wemp.dta	Stata wemp.dta data file	2 days ago
wemp.xlsx	An Excel version of the wemp.dta file	2 days ago
zz_intro_jupyter_20170124_vg_v1.pdf	Introduction to Jupyter	a minute ago

README.md

Jupyter Notebook Repository for Q-Step Edinburgh

An introduction to Jupyter Notebooks - a quick-step towards literate computing and reproducible research

The material in the repository is designed for Edinburgh Q-Step students (http://www.q-step.ed.ac.uk/) who wish to use Jupyter Notebooks.

The material supports a Edinburgh Q-Step Workshop held in January 2017.

Getting Started

The Project Jupyter Website http://jupyter.org/

Jupyter Notebooks a Quick-Step Towards Literate Computing and Reproducible Research

Q – Step Edinburgh, January 2017

Vernon Gayle
University of Edinburgh
@profbigvern