

Geographic Data Science -

Lecture II

Modern Computational

Environments

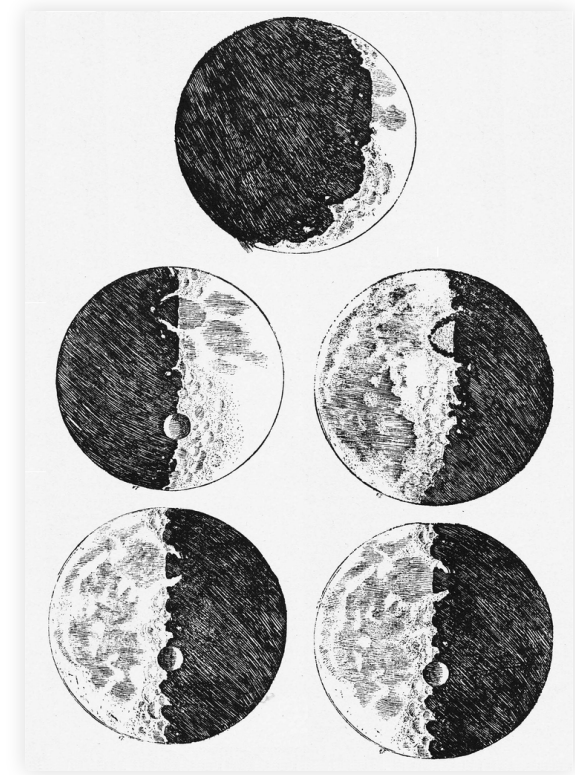
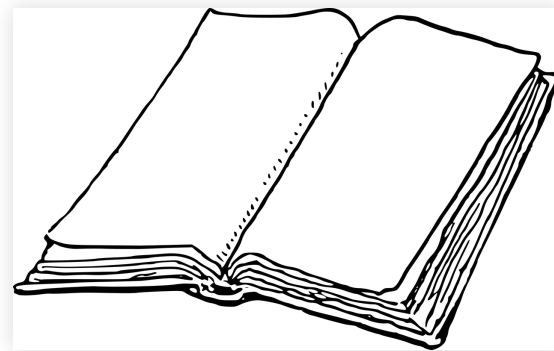
Dani Arribas-Bel

# How does Science “get done”?

- Reproducibility in Science
- Modern scientific tools
- JupyterLab demo

# Reproducibility in Science

# In the old days...

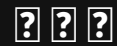




# Reproducibility

- Ability to reproduce scientific procedures (e.g. experiments, results)
- Key to the scientific endeavour
- Embedded in early work

# But...



**Hey! You're not a robot, are you?**

**PRESS & HOLD**

**Access to this page has been denied because we believe you are using automation tools to browse the website**

[Report an issue](#)

Please make sure that Javascript and cookies are enabled on your browser and that you are not blocking them from loading.

[SHARE](#)

Duke University is at the center of a whistleblower lawsuit concerning potential research misconduct. USCHOOLS UNIVERSITY IMAGES/ISTOCKPHOTO

## Whistleblower sues Duke, claims doctored data helped win \$200 million in grants

By [Alison McCook](#), [Retraction Watch](#) | Sep. 1, 2016, 2:00 PM

On a Friday in March 2013, a researcher working in the lab of a prominent pulmonary scientist at Duke University in Durham, North Carolina, was arrested on charges of embezzlement. The researcher, biologist Erin Potts-Kant, later pled guilty to siphoning more than \$25,000 from the Duke University Health System, buying merchandise from Amazon, Walmart, and Target—even faking receipts to legitimize her purchases. A state judge ultimately levied a fine, and sentenced her to probation and community service.

Then Potts-Kant's troubles got worse. Duke officials took a closer look at her work and didn't like what they saw. Fifteen of her papers, mostly dealing with pulmonary biology, have now been retracted, with many notices citing "unreliable" data. Several others have been modified with either partial retractions, expressions of concern, or corrections. And last month, a U.S. district court unsealed a whistleblower lawsuit filed by a former colleague of Potts-Kant. It accuses the researcher, her former supervisor, and the university of including fraudulent data in applications and reports involving more than 60 grants worth some \$200 million. If successful, the suit—brought under the federal False Claims Act (FCA)—could force Duke to return to the government up to three times the amount of any ill-gotten funds, and produce a multimillion-dollar payout to the whistleblower.

The Duke case "should scare all [academic] institutions around the country," says attorney Joel

url

■ April 18, 2013, 11:31 AM GMT+1

# FAQ: Reinhart, Rogoff, and the Excel Error That Changed History

By Peter Coy

url



PHOTOGRAPH BY GREGOR SCHUSTER

Harvard University economists Carmen Reinhart and Kenneth Rogoff have acknowledged making a spreadsheet calculation mistake in a 2010 research paper, “[Growth in a Time of Debt](#)” (PDF), which has been widely cited to justify budget-cutting. But the authors stand by their conclusion that higher government debt is associated with slower economic growth. Here’s what you need to know:



# Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

**Monya Baker**

27 August 2015

 [Rights & Permissions](#)

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

The studies they took on ranged from whether expressing insecurities perpetuates them to differences in how children and adults respond to fear stimuli, to effective ways to teach arithmetic.



Brian Nosek's team set out to replicate scores of studies.

url

# Modern Scientific Tools

# Reproducible Science

- **Transparent** (computational) processes
- Enough **detail** to reproduce the entire analysis
- Efficient model of **reusability** (D.R.Y.)

# Building Blocks

- Computational Notebooks
- Open-Source Packages
- Open Platforms



# Computational Notebooks

url

# The Scientific Paper Is Obsolete

Here's what's next.



PNAS / Richard Goerg / Getty / The Atlantic

JAMES SOMERS | APR 5, 2018 | SCIENCE

Share
Tweet
...

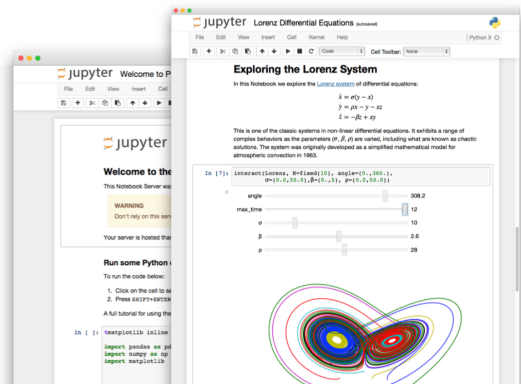
TEXT SIZE

- +

Like *The Atlantic*? Subscribe to [The Atlantic Daily](#), our free weekday email newsletter.

Email  **SIGN UP**

**T**HE SCIENTIFIC paper—the actual form of it—was one of the enabling inventions of modernity. Before it was developed in the 1600s, results were communicated privately in letters, ephemerally in lectures, or all at once in books. There was no public forum for *incremental* advances. By making room for reports of single experiments or minor technical advances, journals made the chaos of science accretive. Scientists from that point forward became like the social insects: They made their progress steadily, as a buzzing mass.



## The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Try it in your browser

Install the Notebook



### Language of choice

Jupyter supports over 40 programming languages, including Python, R, Julia, and Scala.



### Share notebooks

Notebooks can be shared with others using email, Dropbox, GitHub and the [Jupyter Notebook Viewer](#).



### Interactive output

Your code can produce rich, interactive output: HTML, images, videos, LaTeX, and custom MIME types.



### Big data integration

Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, TensorFlow.

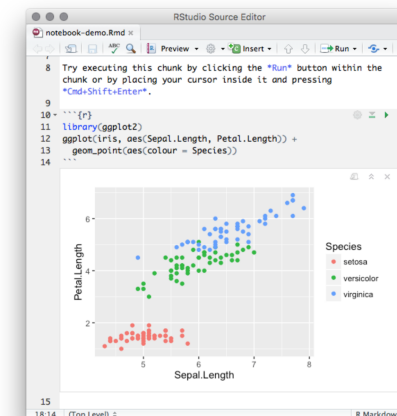
## R Notebooks

Jonathan McPherson

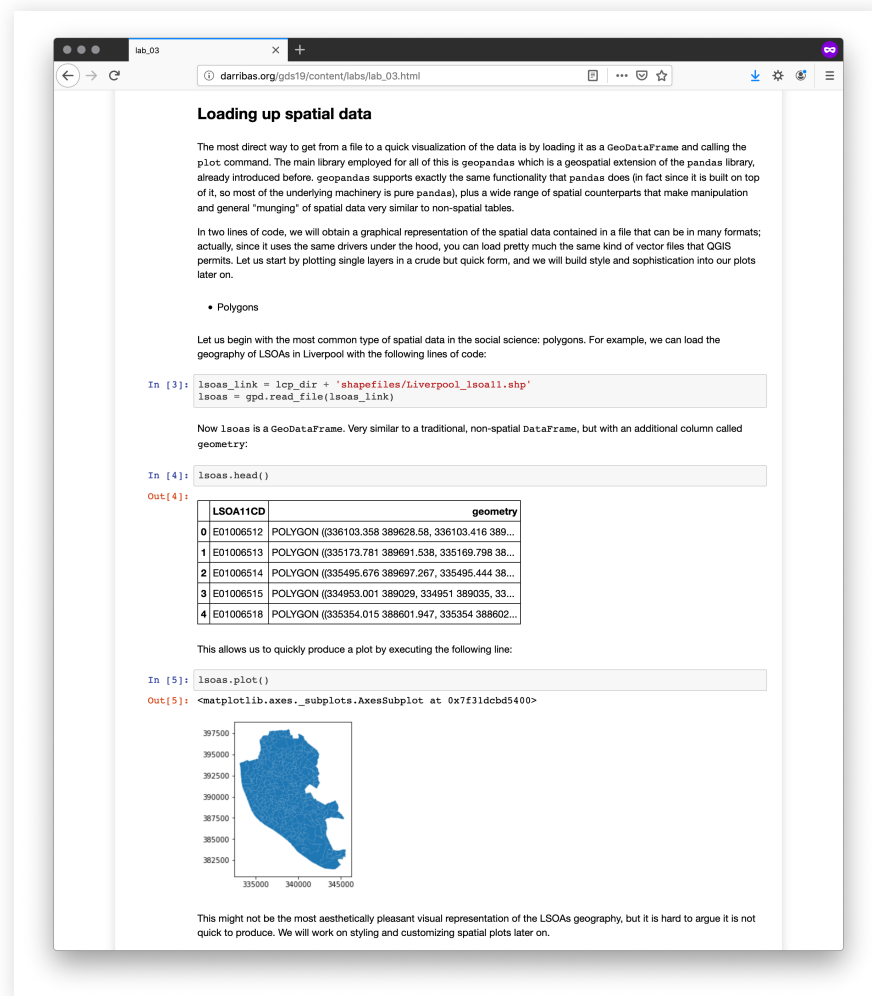
2016-10-05

Categories: [Featured R Markdown](#) [RStudio IDE](#)

Today we're excited to announce [R Notebooks](#), which add a powerful notebook authoring engine to [R Markdown](#). Notebook interfaces for data analysis have compelling advantages including the close association of code and output and the ability to intersperse narrative with computation. Notebooks are also an excellent tool for teaching and a convenient way to share analyses.



# Computational Notebooks



The screenshot shows a Jupyter Notebook interface with the following content:

### Loading up spatial data

The most direct way to get from a file to a quick visualization of the data is by loading it as a `GeoDataFrame` and calling the `plot` command. The main library employed for all of this is `geopandas` which is a geospatial extension of the `pandas` library, already introduced before. `geopandas` supports exactly the same functionality that `pandas` does (in fact since it is built on top of it, so most of the underlying machinery is pure `pandas`), plus a wide range of spatial counterparts that make manipulation and general "munging" of spatial data very similar to non-spatial tables.

In two lines of code, we will obtain a graphical representation of the spatial data contained in a file that can be in many formats; actually, since it uses the same drivers under the hood, you can load pretty much the same kind of vector files that QGIS permits. Let us start by plotting single layers in a crude but quick form, and we will build style and sophistication into our plots later on.

- Polygons

Let us begin with the most common type of spatial data in the social science: polygons. For example, we can load the geography of LSOAs in Liverpool with the following lines of code:

```
In [3]: lsoas_link = lcp_dir + 'shapefiles/Liverpool_lsoa11.shp'
lsoas = gpd.read_file(lsoas_link)
```

Now `lsoas` is a `GeoDataFrame`. Very similar to a traditional, non-spatial `DataFrame`, but with an additional column called `geometry`:

```
In [4]: lsoas.head()
```

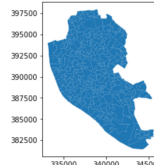
```
Out[4]:
```

	LSOA11CD	geometry
0	E01006512	POLYGON ((336103.358 389628.56, 336103.416 389...
1	E01006513	POLYGON ((335173.781 388691.538, 335169.798 38...
2	E01006514	POLYGON ((335495.676 389697.267, 335495.444 38...
3	E01006515	POLYGON ((334953.001 389029, 334951 389035, 33...
4	E01006518	POLYGON ((335354.015 388601.947, 335354 388602...

This allows us to quickly produce a plot by executing the following line:

```
In [5]: lsoas.plot()
```

```
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f31dcbd5400>
```



This might not be the most aesthetically pleasant visual representation of the LSOAs geography, but it is hard to argue it is not quick to produce. We will work on styling and customizing spatial plots later on.

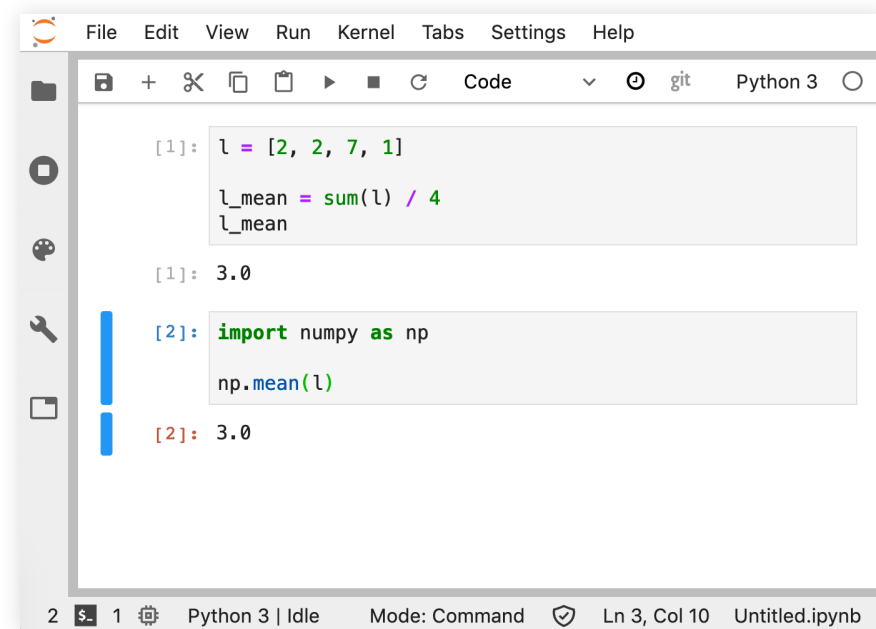
One-file documents:

- (Executable) code
- Output
- Narrative text

# Open-Source Packages

D.R.Y. (Don't Repeat Yourself)

- Encapsulate reusable functionality
- Easy access + more reliable (if package is good!)
- Code available (free as in beer... but also as in speech!)



```
File Edit View Run Kernel Tabs Settings Help
+ ✂ 📄 ▶ ■ 🔄 Code git Python 3
[1]: l = [2, 2, 7, 1]
      l_mean = sum(l) / 4
      l_mean
[1]: 3.0
[2]: import numpy as np
      np.mean(l)
[2]: 3.0
2 Python 3 | Idle Mode: Command Ln 3, Col 10 Untitled.ipynb
```

# Platforms

*Hardware and low-level software (OS) that supports computations*

Change of models:

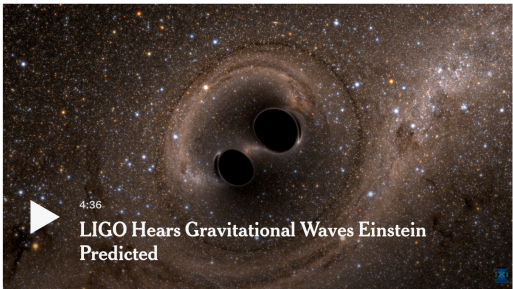
- Desktop Vs cloud
- Integrated Vs distributed
- Native installation Vs  
virtualisation/containerisation

# Examples

# Ligo gravitational waves

SCIENCE **The New York Times** LOG IN

## 2017 Nobel Prize in Physics Awarded to LIGO Black Hole Researchers



4:36  
LIGO Hears Gravitational Waves Einstein Predicted

About a hundred years ago, Einstein predicted the existence of gravitational waves, but until now, they were undetectable. Artist's rendering/Simulating eXtreme Spacetimes

By Dennis Overbye  
Oct. 3, 2017

Facebook Twitter Email Print Bookmark

GW Open Science Center  
https://www.gw-openscience.org/tutorials/

## Gravitational Wave Open Science Center

Home Data Software Online Status About GWOSC

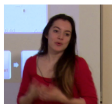
### Tutorials

Each tutorial will lead you step-by-step through some common data analysis tasks. While GWOSC data can be analyzed using libraries in many software languages (C, C++, Matlab, etc.), most of these tutorials use Python. See also the [software page](#) for more examples.

See the [tutorial setup page](#) for help installing software to run these tutorials.

Tutorials shown here are not used to produce published results. For gravitational-wave software analysis packages that are used to produce LSC and Virgo Collaboration publications, see [software page](#).

### Gravitational Wave Open Data Workshop Web Course (2019)



Lecture videos and tutorials from 2019 workshop  
[Course Material](#)



# 2019



**The Spectator Index**  
@spectatorindex

IMAGE: Reaction of Katie Bouman, who led the creation of an algorithm to produce first image of black hole.

A photograph of Katie Bouman in a computer lab, reacting with her hands to her mouth in a surprised or excited expression. A laptop in the foreground shows the first image of a black hole.

1:15 AM · Apr 11, 2019 · [TweetDeck](#)

2.5K Retweets 8.5K Likes

**NATIONAL GEOGRAPHIC**

The first image of a supermassive black hole, showing a bright ring of light surrounding a dark central shadow.

The Event Horizon Telescope—a planet-scale array of ground-based radio telescopes—has obtained the first image of a supermassive black hole and its shadow. The image reveals the central black hole of Messier 87, a massive galaxy in the Virgo cluster.

PHOTOGRAPH BY EVENT HORIZON TELESCOPE COLLABORATION

SCIENCE & INNOVATION | STARSTRUCK

### First-ever picture of a black hole unveiled

Using a telescope the size of the planet, astronomers have captured the first image of this space oddity. Here's why that matters.

9 MINUTE READ

BY [NADIA DRAKE](#)

PUBLISHED APRIL 10, 2019

JupyterLab (live) demo



Geographic Data Science'19 by Dani Arribas-Bel is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.