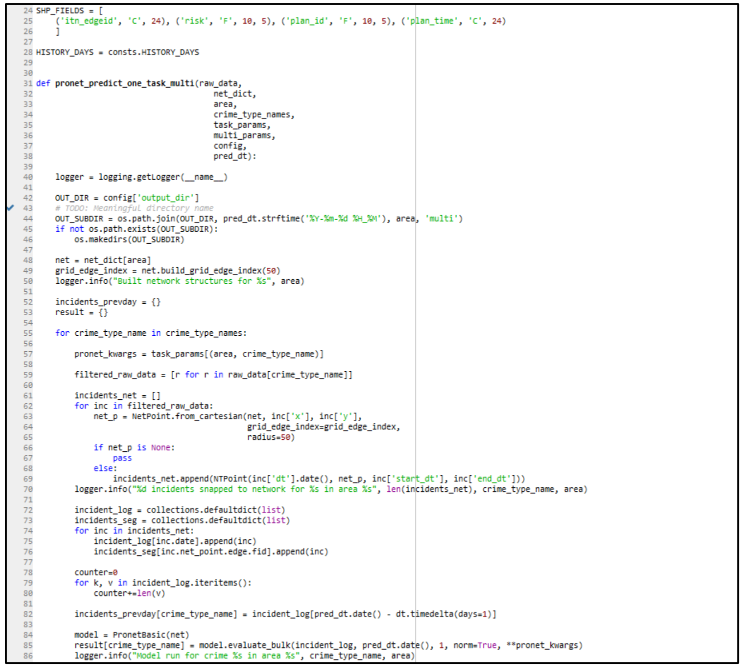
Code Sharing

JDI OPEN

Toby Davies

What is code sharing?

Code sharing is making available – publicly – the code that you use to analyse your data

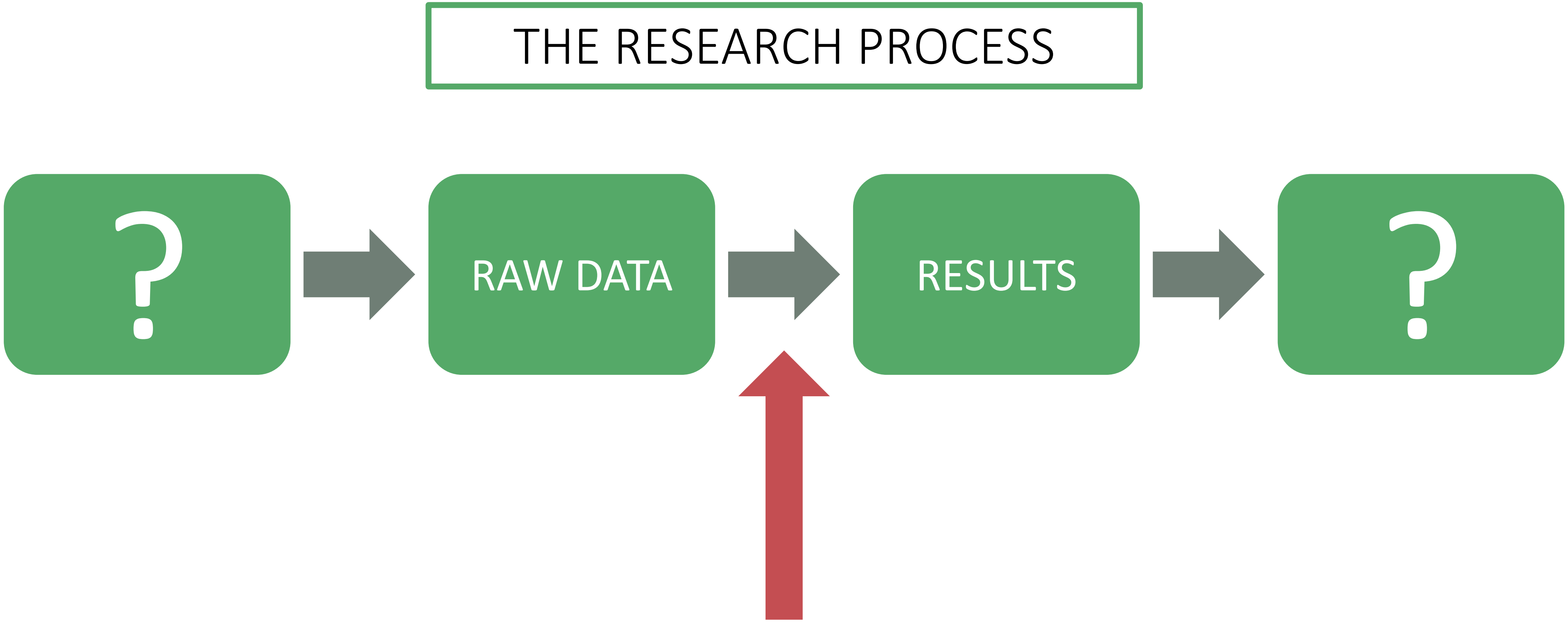


```
SHP_FIELDS = [
    ('itn_edgeid', 'C', 24), ('risk', 'F', 10, 5), ('plan_id', 'F', 10, 5), ('plan_time', 'C', 24)
    ]

HISTORY_DAYS = consts.HISTORY_DAYS


def pronet_predict_one_task_multi(raw_data,
                                     net_dict,
                                     area,
                                     crime_type_names,
                                     task_params,
                                     multi_params,
                                     config,
                                     pred_dt):

    logger = logging.getLogger(__name__)

    OUT_DIR = config['output_dir']
    # TODO: Meaningful directory name
    OUT_SUBDIR = os.path.join(OUT_DIR, pred_dt.strftime('%Y-%m-%d %H_%M'), area, 'multi')
    if not os.path.exists(OUT_SUBDIR):
        os.makedirs(OUT_SUBDIR)

    net = net_dict[area]
    grid_edge_index = net.build_grid_edge_index(50)
    logger.info("Built network structures for %s", area)

    incidents_prevday = {}
    result = {}

    for crime_type_name in crime_type_names:

        pronet_kwargs = task_params[(area, crime_type_name)]

        filtered_raw_data = [r for r in raw_data[crime_type_name]]

        incidents_net = []
        for inc in filtered_raw_data:
            net_p = NetPoint.from_cartesian(net, inc['x'], inc['y'],
                                            grid_edge_index=grid_edge_index,
                                            radius=50)
            if net_p is None:
                pass
            else:
                incidents_net.append(NTPoint(inc['dt'].date(), net_p, inc['start_dt'], inc['end_dt']))
        logger.info("%d incidents snapped to network for %s in area %s", len(incidents_net), crime_type_name, area)

        incident_log = collections.defaultdict(list)
        incidents_seg = collections.defaultdict(list)
        for inc in incidents_net:
            incident_log[inc.date].append(inc)
            incidents_seg[inc.net_point.edge.fid].append(inc)

        counter=0
        for k, v in incident_log.iteritems():
            counter+=len(v)

        incidents_prevday[crime_type_name] = incident_log[pred_dt.date() - dt.timedelta(days=1)]

        model = PronetBasic(net)
        result[crime_type_name] = model.evaluate_bulk(incident_log, pred_dt.date(), 1, norm=True, **pronet_kwargs)
        logger.info("Model run for crime %s in area %s", crime_type_name, area)
```

…and that code existing in the first place.

This talk is really about **reproducible** analysis more generally…

What do you mean ‘code’?



THE RESEARCH PROCESS

?

RAW DATA

RESULTS

?

‘CODE’

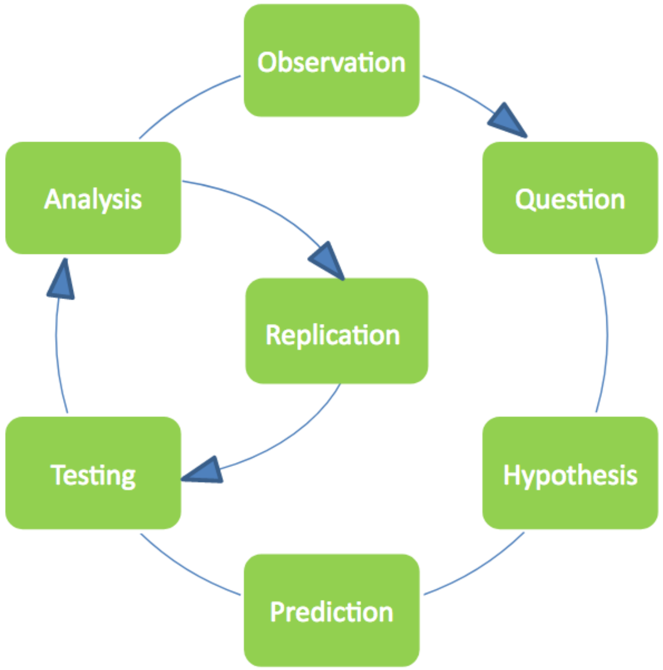
(or something equivalent)

Why share?

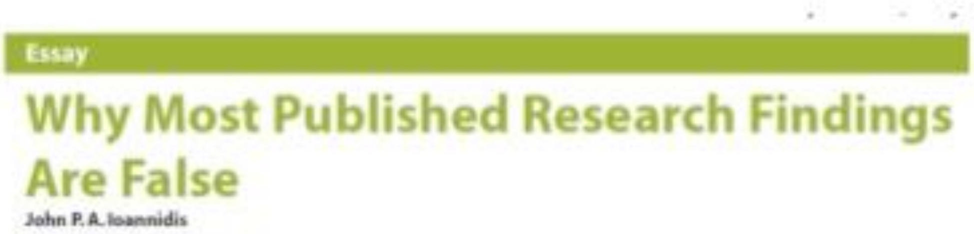
Reproducibility

We all already know…

- Replication and reproduction is at the heart of the scientific method
- Findings are only credible if they reproduce – highest standard of evidence
- Re-use of methods allows extension of knowledge



Reproducibility Crisis

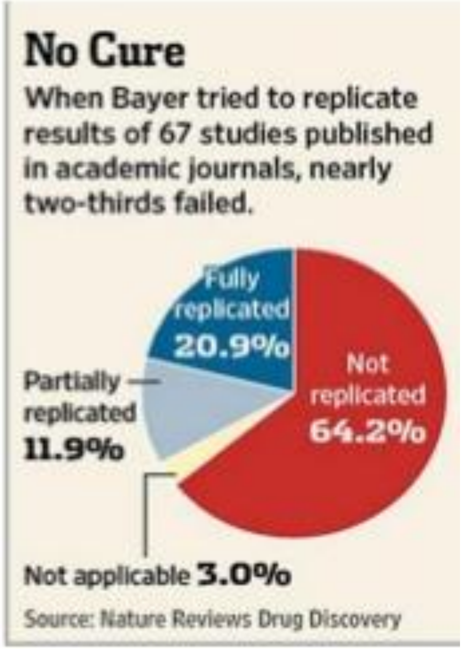


Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

2005. PLoS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124



No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Source: Nature Reviews Drug Discovery

"There is increasing concern about the reliability of biomedical research, with recent articles suggesting that up to 85% of research funding is wasted."

Bustin, S. A. (2015). The reproducibility of biomedical research: Sleepers awake! *Biomolecular Detection and Quantification*

THE LANCET

Research: increasing value, reducing waste

Published: January 8, 2014

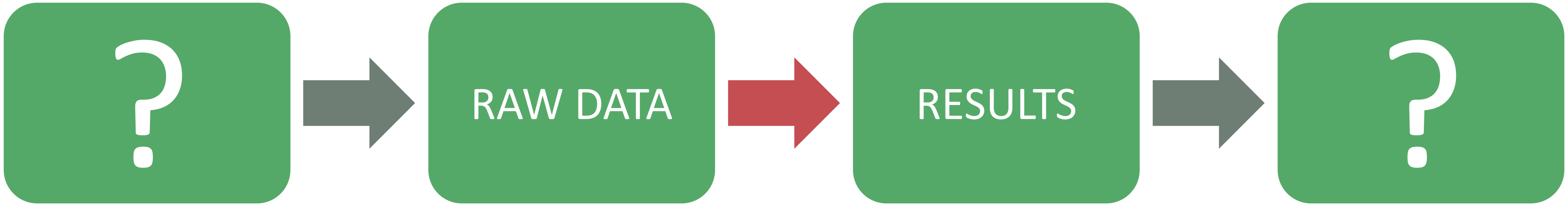
NATURE | NEWS

First results from psychology's largest reproducibility test

Reproducibility Crisis

We've discussed the (probable) reasons for this:

Genuine Mistakes	Questionable Research Practices	Wilful Misconduct



?

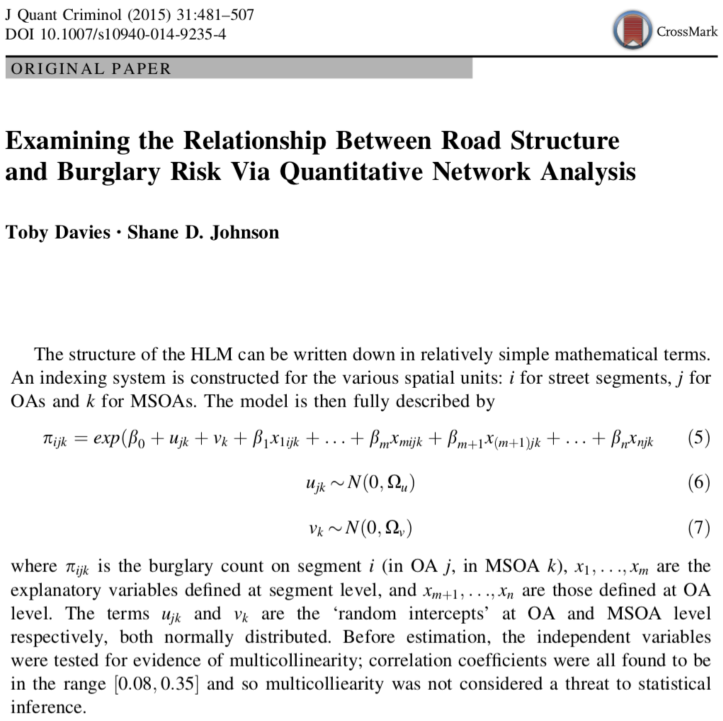
RAW DATA

RESULTS

?

Don't even need a crisis for these to be problems…

Reporting



J Quant Criminol (2015) 31:481–507
DOI 10.1007/s10940-014-9235-4

ORIGINAL PAPER

Examining the Relationship Between Road Structure and Burglary Risk Via Quantitative Network Analysis

Toby Davies · Shane D. Johnson

The structure of the HLM can be written down in relatively simple mathematical terms. An indexing system is constructed for the various spatial units: i for street segments, j for OAs and k for MSOAs. The model is then fully described by

$$\pi_{ijk} = exp(\beta_0 + u_{jk} + v_k + \beta_1 x_{1ijk} + \ldots + \beta_m x_{mijk} + \beta_{m+1} x_{(m+1)jk} + \ldots + \beta_n x_{njk} \qquad (5)$$

$$u_{jk} \sim N(0, \Omega_u) \qquad (6)$$

$$v_k \sim N(0, \Omega_v) \qquad (7)$$

where π_{ijk} is the burglary count on segment i (in OA j, in MSOA k), $x_1, \ldots, x_m$ are the explanatory variables defined at segment level, and $x_{m+1}, \ldots, x_n$ are those defined at OA level. The terms u_{jk} and v_k are the 'random intercepts' at OA and MSOA level respectively, both normally distributed. Before estimation, the independent variables were tested for evidence of multicollinearity; correlation coefficients were all found to be in the range $[0.08, 0.35]$ and so multicolliearity was not considered a threat to statistical inference.

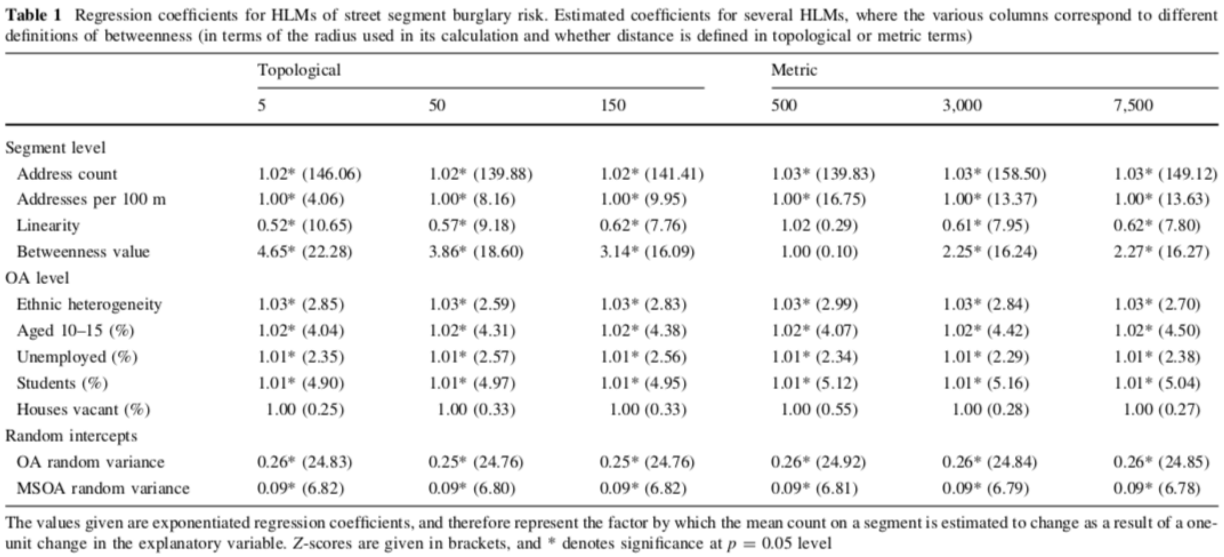


Table 1 Regression coefficients for HLMs of street segment burglary risk. Estimated coefficients for several HLMs, where the various columns correspond to different definitions of betweenness (in terms of the radius used in its calculation and whether distance is defined in topological or metric terms)

	Topological			Metric		
	5	50	150	500	3,000	7,500
Segment level						
Address count	1.02* (146.06)	1.02* (139.88)	1.02* (141.41)	1.03* (139.83)	1.03* (158.50)	1.03* (149.12)
Addresses per 100 m	1.00* (4.06)	1.00* (8.16)	1.00* (9.95)	1.00* (16.75)	1.00* (13.37)	1.00* (13.63)
Linearity	0.52* (10.65)	0.57* (9.18)	0.62* (7.76)	1.02 (0.29)	0.61* (7.95)	0.62* (7.80)
Betweenness value	4.65* (22.28)	3.86* (18.60)	3.14* (16.09)	1.00 (0.10)	2.25* (16.24)	2.27* (16.27)
OA level						
Ethnic heterogeneity	1.03* (2.85)	1.03* (2.59)	1.03* (2.83)	1.03* (2.99)	1.03* (2.84)	1.03* (2.70)
Aged 10–15 (%)	1.02* (4.04)	1.02* (4.31)	1.02* (4.38)	1.02* (4.07)	1.02* (4.42)	1.02* (4.50)
Unemployed (%)	1.01* (2.35)	1.01* (2.57)	1.01* (2.56)	1.01* (2.34)	1.01* (2.29)	1.01* (2.38)
Students (%)	1.01* (4.90)	1.01* (4.97)	1.01* (4.95)	1.01* (5.12)	1.01* (5.16)	1.01* (5.04)
Houses vacant (%)	1.00 (0.25)	1.00 (0.33)	1.00 (0.33)	1.00 (0.55)	1.00 (0.28)	1.00 (0.27)
Random intercepts						
OA random variance	0.26* (24.83)	0.25* (24.76)	0.25* (24.76)	0.26* (24.92)	0.26* (24.84)	0.26* (24.85)
MSOA random variance	0.09* (6.82)	0.09* (6.80)	0.09* (6.82)	0.09* (6.81)	0.09* (6.79)	0.09* (6.78)

The values given are exponentiated regression coefficients, and therefore represent the factor by which the mean count on a segment is estimated to change as a result of a one-unit change in the explanatory variable. *Z*-scores are given in brackets, and * denotes significance at $p = 0.05$ level

Even if I do everything right, I'm asking you to believe:

- I did what I claimed I did
- …and nothing else
- I implemented it correctly
- The numbers are what came out of the software

Reporting

Good reasons for this kind of reporting…

…in the 19th century

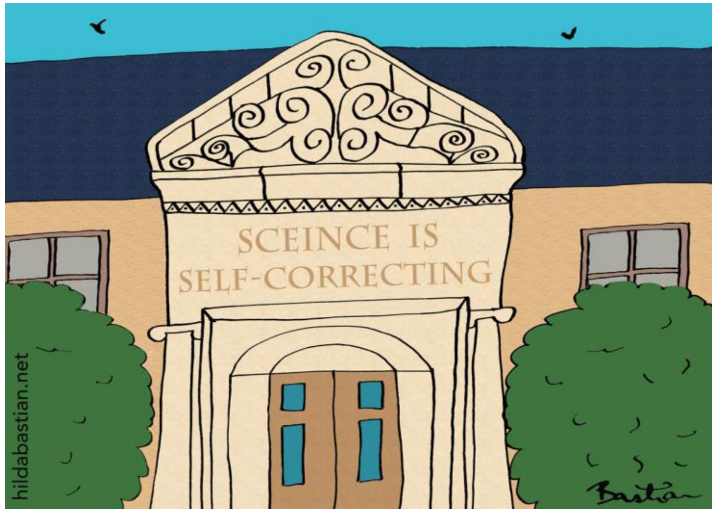
2018

2018

Scientific rigour

So being as open as possible with methods is good because it promotes:

- **Analytic reproducibility** – whether the analysis ‘works’; checking for errors in the analysis pipeline
- **Analytic robustness** – whether conclusions are still valid under alternative specifications
- **Replicability** – the application of the same procedures to alternative data



Why else?

More generally, greater transparency is good for the field:

- Allows others to adapt and build on your work
- …and do it ‘properly’
- Makes scientific discovery more efficient
- Promotes (genuine) replication – particularly important in a field where data sharing is challenging
- Ethically ‘right’ – someone paid for the work

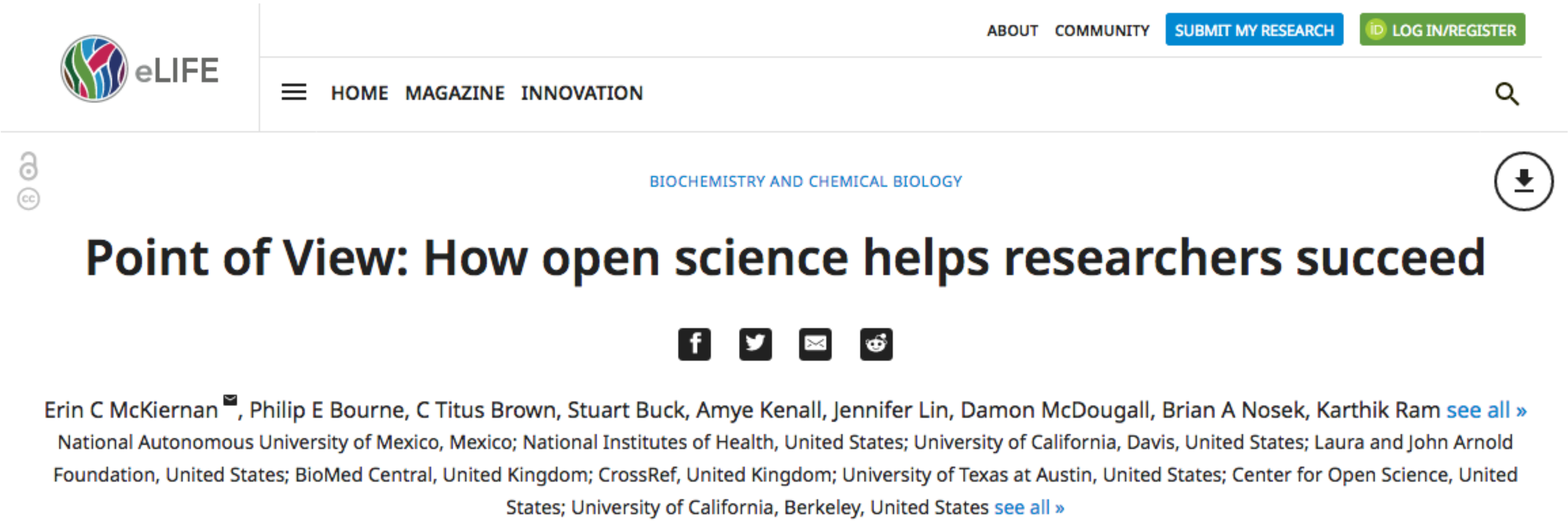
More selfishly…

The things that constitute good reproducible practice are exactly those that help your own work.

> ‘Your closest collaborator is you six months ago, but you don’t reply to emails’
>
> *Broman (2016)*

Kudos

People will think you are cool.



BIOCHEMISTRY AND CHEMICAL BIOLOGY

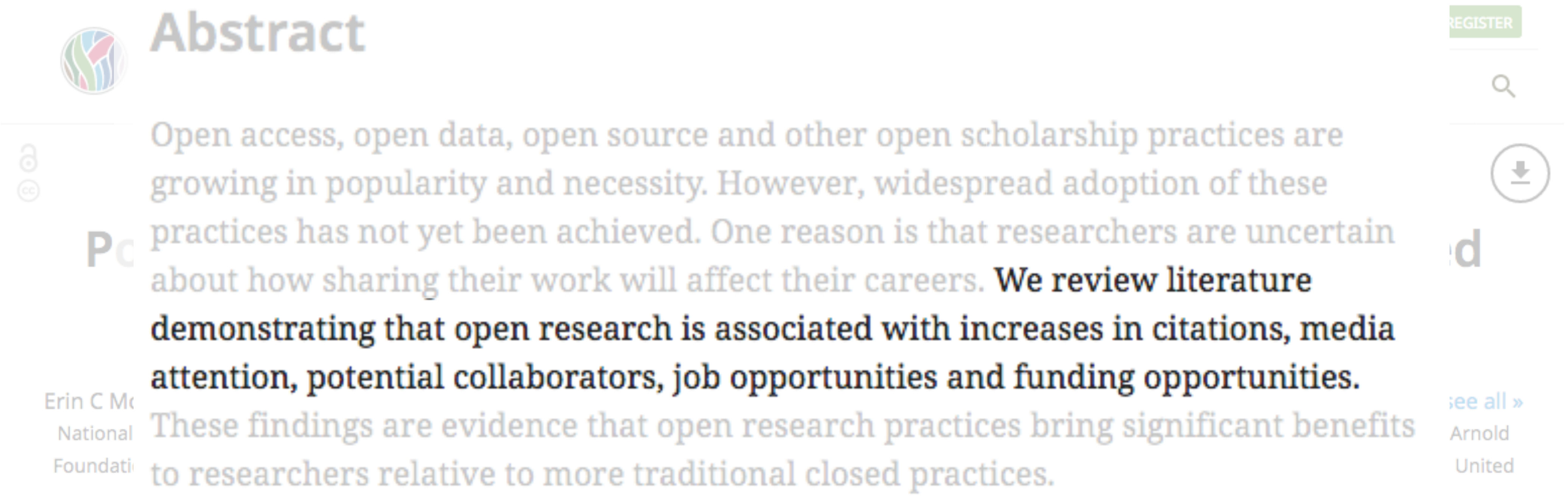
Point of View: How open science helps researchers succeed

Erin C McKiernan, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram see all »

National Autonomous University of Mexico, Mexico; National Institutes of Health, United States; University of California, Davis, United States; Laura and John Arnold Foundation, United States; BioMed Central, United Kingdom; CrossRef, United Kingdom; University of Texas at Austin, United States; Center for Open Science, United States; University of California, Berkeley, United States see all »

Kudos

People will think you are cool.



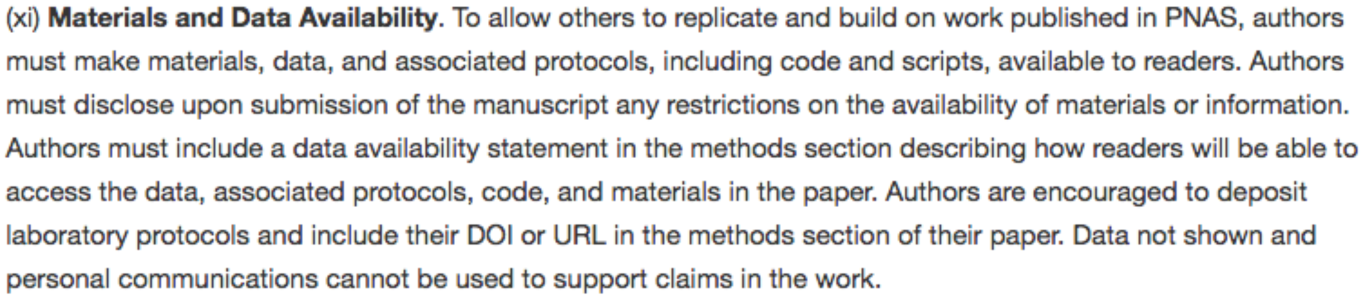
Abstract

Open access, open data, open source and other open scholarship practices are growing in popularity and necessity. However, widespread adoption of these practices has not yet been achieved. One reason is that researchers are uncertain about how sharing their work will affect their careers. **We review literature demonstrating that open research is associated with increases in citations, media attention, potential collaborators, job opportunities and funding opportunities.** These findings are evidence that open research practices bring significant benefits to researchers relative to more traditional closed practices.

The stick

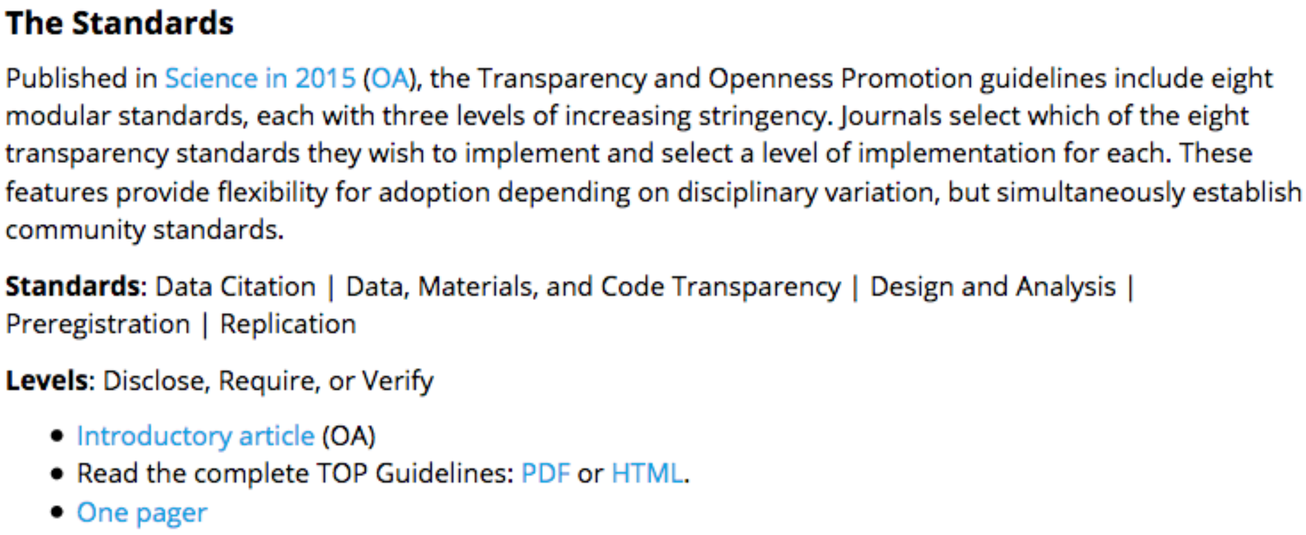
As crime scientists, we know that the best way to promote good behaviour is via threats; the more punitive the better.

PNAS Proceedings of the National Academy of Sciences of the United States of America



(xi) **Materials and Data Availability**. To allow others to replicate and build on work published in PNAS, authors must make materials, data, and associated protocols, including code and scripts, available to readers. Authors must disclose upon submission of the manuscript any restrictions on the availability of materials or information. Authors must include a data availability statement in the methods section describing how readers will be able to access the data, associated protocols, code, and materials in the paper. Authors are encouraged to deposit laboratory protocols and include their DOI or URL in the methods section of their paper. Data not shown and personal communications cannot be used to support claims in the work.

The stick



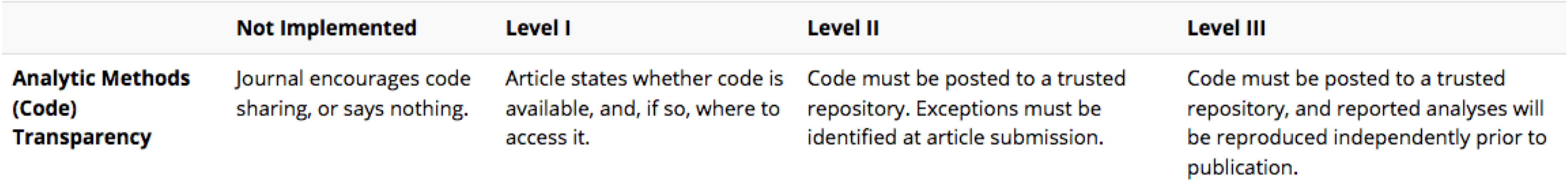
The Standards

Published in Science in 2015 (OA), the Transparency and Openness Promotion guidelines include eight modular standards, each with three levels of increasing stringency. Journals select which of the eight transparency standards they wish to implement and select a level of implementation for each. These features provide flexibility for adoption depending on disciplinary variation, but simultaneously establish community standards.

Standards: Data Citation | Data, Materials, and Code Transparency | Design and Analysis | Preregistration | Replication

Levels: Disclose, Require, or Verify

- Introductory article (OA)
- Read the complete TOP Guidelines: PDF or HTML.
- One pager



	Not Implemented	Level I	Level II	Level III
Analytic Methods (Code) Transparency	Journal encourages code sharing, or says nothing.	Article states whether code is available, and, if so, where to access it.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.

A step back

Before we get to code **sharing**, there is a lot of best practice to cover with respect to reproducibility itself…

‘Lab book’

As a minimum, records should be kept of EXACTLY how any particular result was generated:

- Many analysis workflows involve many tools
- Even though the key step may be contained in only one, pre- and post-processing is critical in order to reach the achieved result
- Should cover the full journey from raw data (and I mean raw) to result
- Name and version of software, along with all parameters used
- Workflow management systems provide a means to store these…
- …but at a minimum the notes must exist.

Avoid manual manipulation

Almost every workflow involves some data cleaning or wrangling. But ‘manual’ approaches to this (point-and-click, copy-paste, delete) are problematic:

- They are error-prone
- They are ambiguous and can be difficult to reproduce

This can all be avoided by **scripting** analysis – manipulating data using specified written commands:

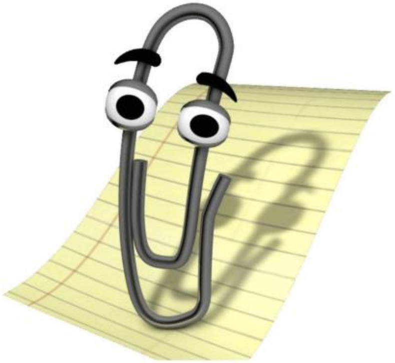
- This often means adopting a programming approach such as R or Python
- …but it doesn’t have to
- This seems like a prohibitive learning curve, but it pays off many times over
- …and makes things more accessible

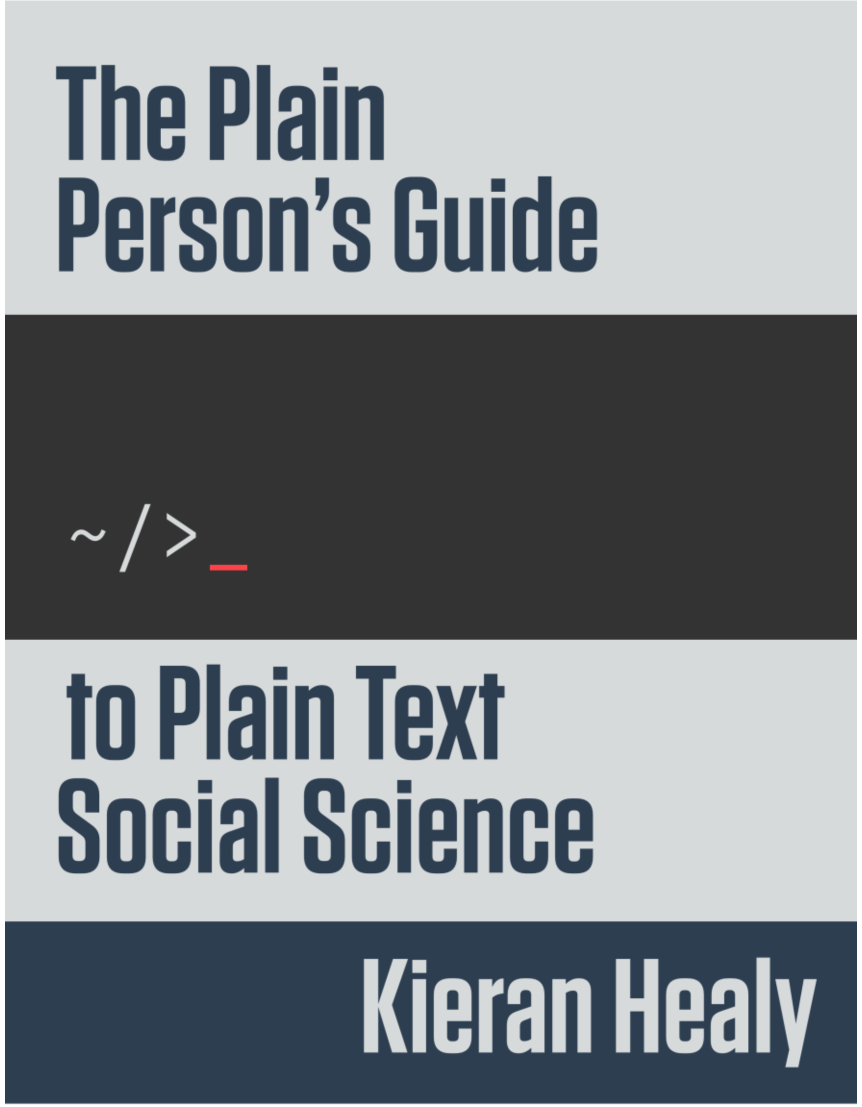
Use ‘simple’ formats

The best way to promote inter-operability between programmes and users is to store data, code and text in simple, universal formats.

- Plain text (.txt)
- Comma separated variables (.csv)

People can open these; they can’t necessarily open proprietary or idiosyncratic file types. Transparency isn’t transparency if not easily accessible…





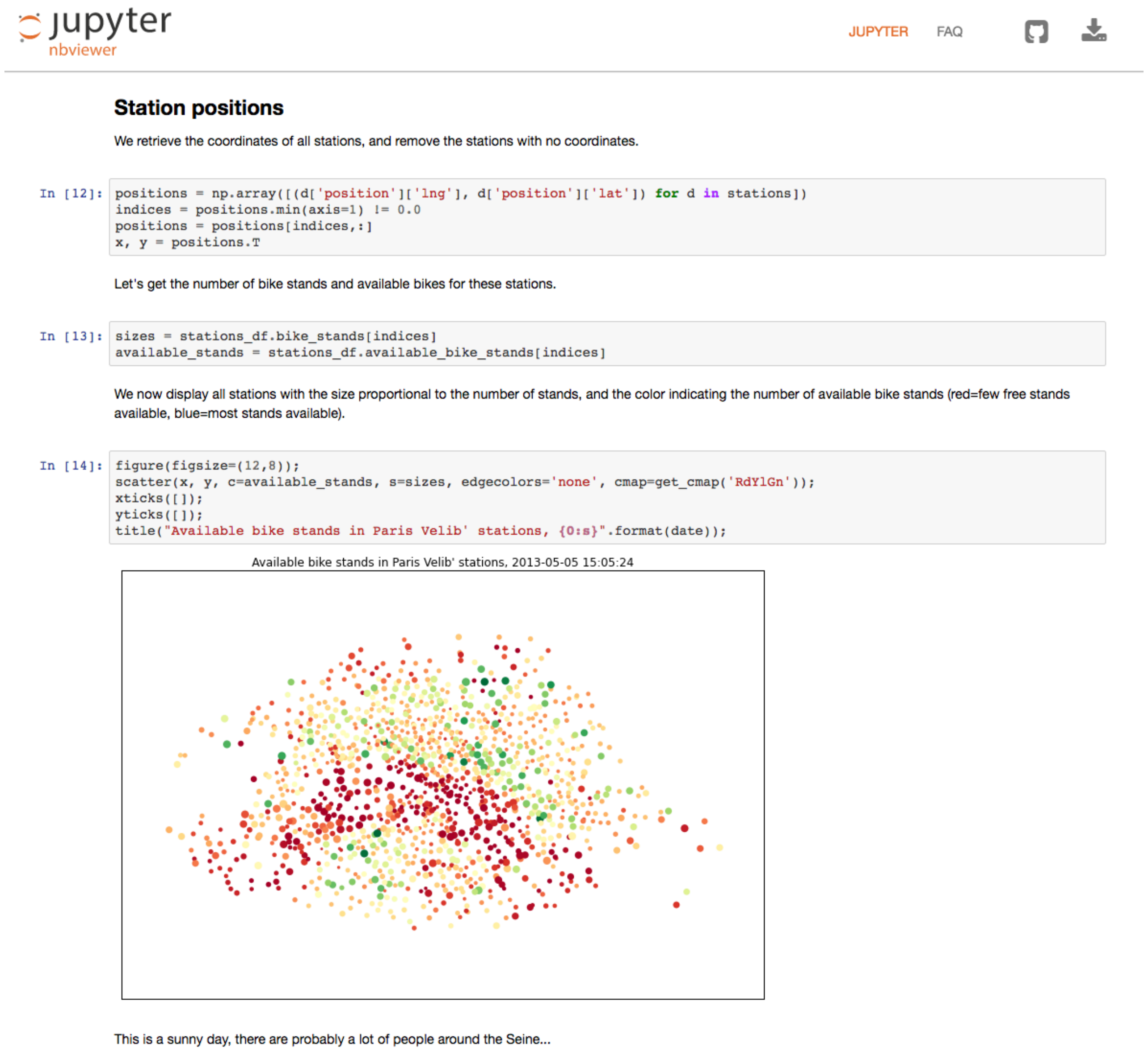
The Jupyter notebook

Provides a way to produce ‘documents with code’…

Why is this good?

- Can seamlessly weave together analysis and narrative
- Code is broken down into small chunks, doing small tasks
- Completely transparent
- Easy to communicate
- **REPLICATION**

Equivalent tools available for R and others…



Station positions

We retrieve the coordinates of all stations, and remove the stations with no coordinates.

```
In [12]: positions = np.array([(d['position']['lng'], d['position']['lat']) for d in stations])
         indices = positions.min(axis=1) != 0.0
         positions = positions[indices,:]
         x, y = positions.T
```

Let's get the number of bike stands and available bikes for these stations.

```
In [13]: sizes = stations_df.bike_stands[indices]
         available_stands = stations_df.available_bike_stands[indices]
```

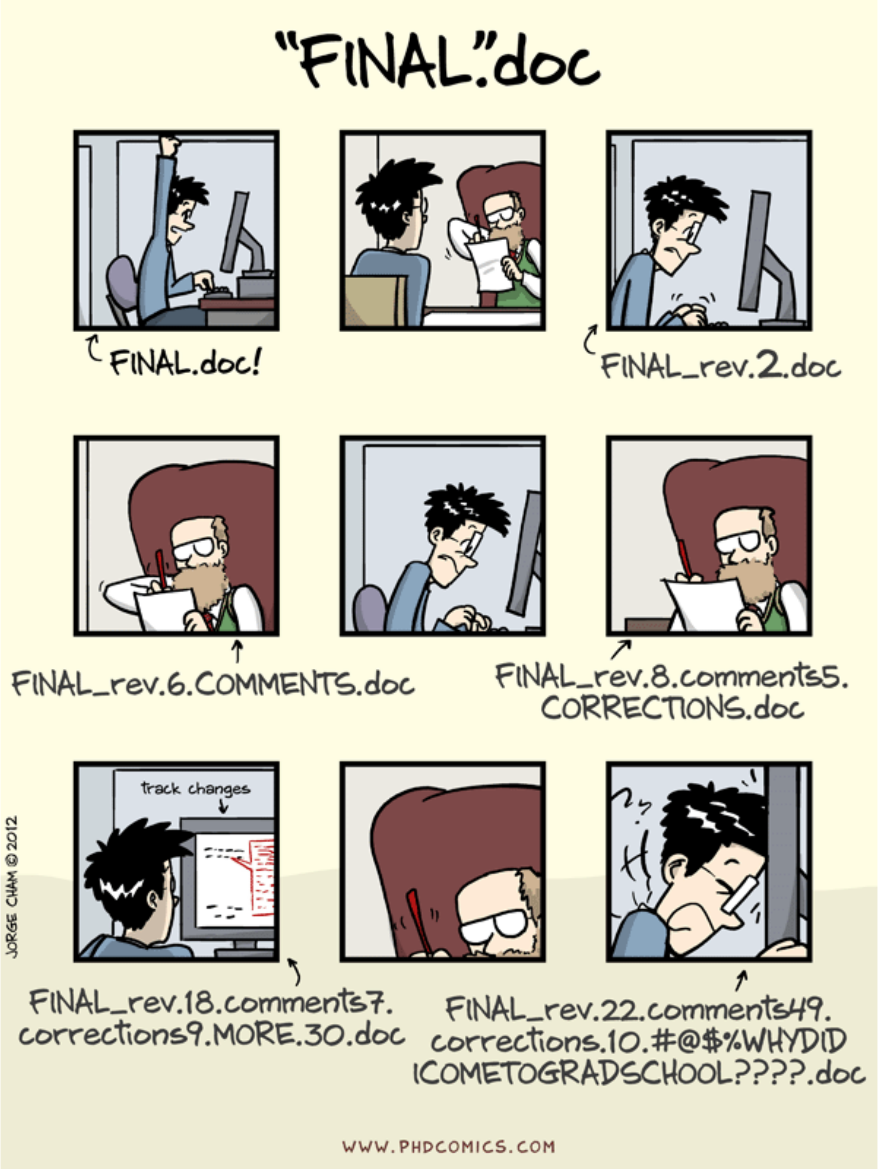
We now display all stations with the size proportional to the number of stands, and the color indicating the number of available bike stands (red=few free stands available, blue=most stands available).

```
In [14]: figure(figsize=(12,8));
         scatter(x, y, c=available_stands, s=sizes, edgecolors='none', cmap=get_cmap('RdYlGn'));
         xticks([]);
         yticks([]);
         title("Available bike stands in Paris Velib' stations, {0:s}".format(date));
```

Available bike stands in Paris Velib' stations, 2013-05-05 15:05:24

This is a sunny day, there are probably a lot of people around the Seine...

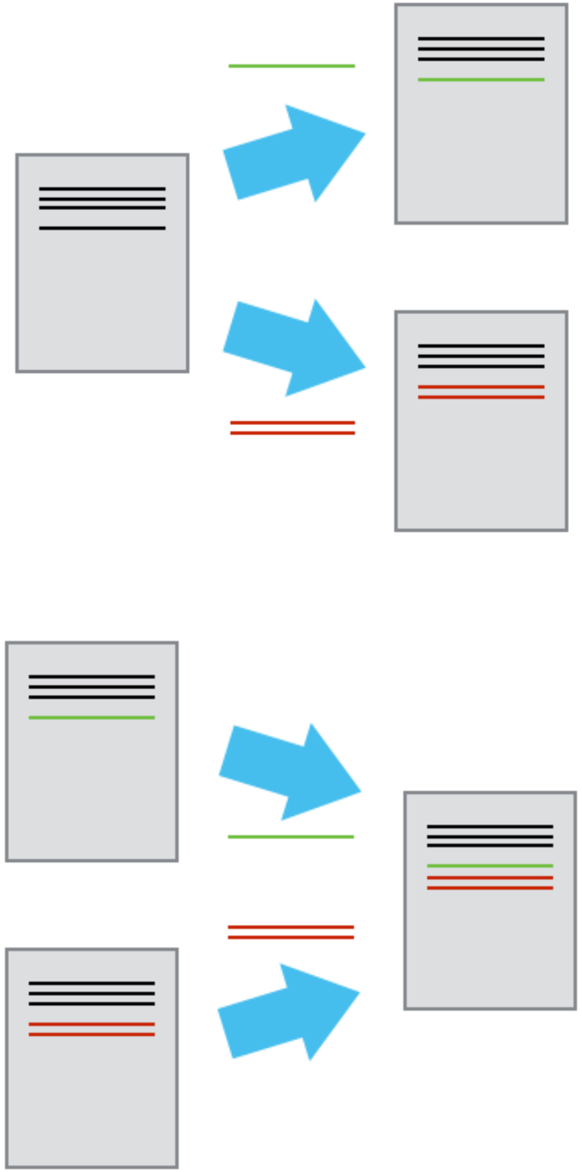
Version control



Version control

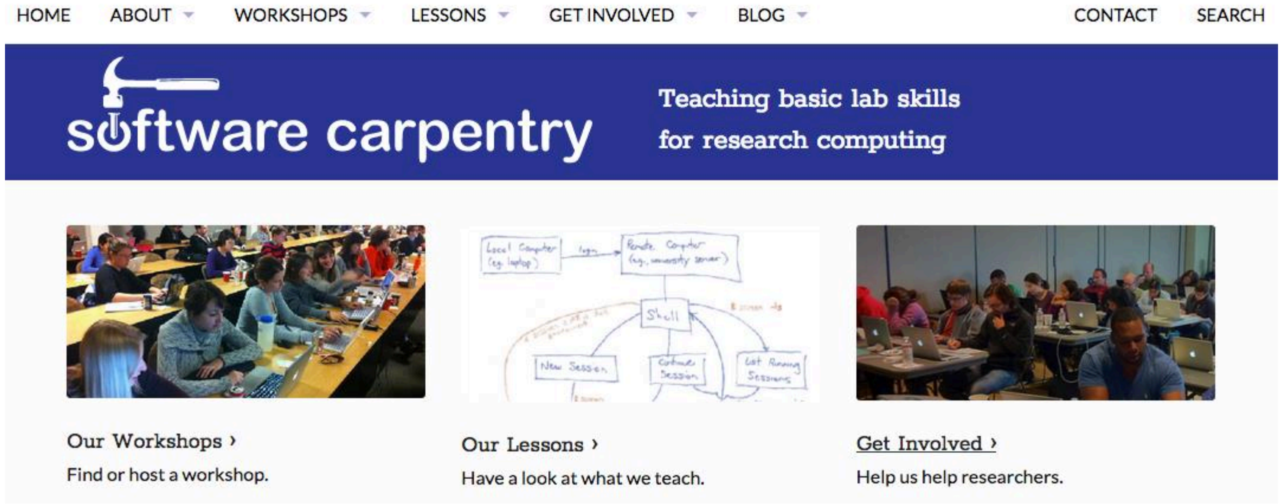
Version control systems end this forever:

- One single version of every file
- Tracks changes at every stage, with a fully documented log
- Like having an unlimited ‘undo’ and ability to roll back to any point – such as the point before you ruined your document
- Facilitates collaboration by tracking everyone’s contributions
- Ability to create branches – parallel worlds – to experiment with new things without breaking the main work
- Not just for code – manuscripts, websites…
- Worth the effort x1000



Github

The Carpentries



HOME ABOUT WORKSHOPS LESSONS GET INVOLVED BLOG CONTACT SEARCH

software carpentry

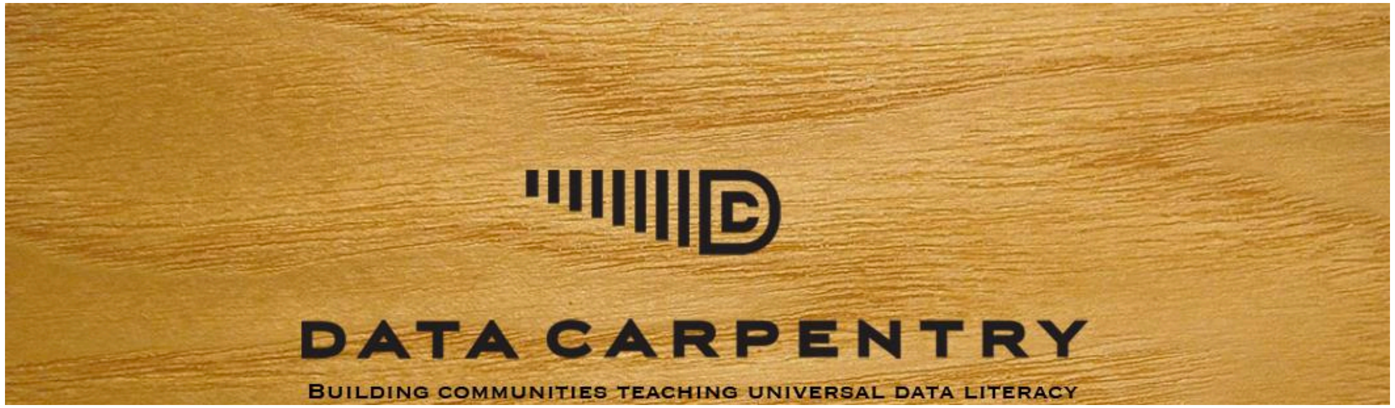
Teaching basic lab skills
for research computing

Local Computer (e.g. laptop)
login
Remote Computer (e.g. university server)
Shell
New Session
Continue Session
List Running Sessions

Our Workshops ›
Find or host a workshop.

Our Lessons ›
Have a look at what we teach.

Get Involved ›
Help us help researchers.



DATA CARPENTRY

Building communities teaching universal data literacy

Sharing code

If your code is on Github, you’ve already done it…

- ‘Code for this study can be found at XXX’

If not, plenty of other places to store:

- OSF, BitBucket, other online repositories

Why not?

You’re still not convinced…

‘I’m worried that I will be scooped’

- Puh-lease…
- Nowhere near as common as people think
- No better proof of precedence than a full Github commit trail…

‘I’m worried that people will find errors in my work’

- So am I
- This is healthy – would you rather just not know?
- So much more sympathy if people are open

‘This will take a lot of time to implement’

- Indeed it will
- But it will also make you a lot more productive