Open Science Working Group Meeting 18/7/19

Science is on a journey...

The Past:
Replication Crisis

The Future:
Reproducibility

Replication Issues in Science

- Ionnidis (2005), *PLOS Medicine*, most published research findings are false.
- Prinz et al. (2011), *Nature Reviews Drug Discovery*, around 65% of cancer biology studies do not replicate.
- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience.
- MacLeod et al. (2014), *Lancet*, 85% of biomedical research resources are wasted.
- Baker (2015), *Nature*, 90% of scientists recognise a ‘reproducibility crisis’.
- Nosek & Errington (2017), *eLife*, out of first 5 replication attempts of preclinical cancer biology work, only 2 have replicated.
- Eisner (2018), *Journal of Molecular and Cellular Cardiology*. Reproducibility of science: Fraud, impact factors and carelessness.

Why are so many studies not replicating?

- There are too many studies with experimental power too low to detect the effect size of interest.

- One of the consequences of a low powered study is that when real effects are detected their magnitude is likely to be over-estimated.

- Studies which find the effect are published and studies that don’t are not published - due to a bias to publish positive results.

- Future work may use the published effect size during *a priori* power analysis (and then fail to find the effect as the new study is effectively under-powered for what it’s looking for).

- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience. Nord et al., (2017), *Journal of Neuroscience,* highlight wide heterogeneity in power in neuroscience studies.

Table 2. Median, maximum, and minimum power subdivided by study type

Group	Median power (%)	Minimum power (%)	Maximum power (%)	2.5th and 97.5th percentile (based on raw data)	95% HDI (based on GMMs)	Total N
All studies	23	0.05	1	0.05–1.00	0.00 – 0.72, 0.80 – 1.00	730
All studies excluding null	30	0.05	1	0.05–1.00	0.01 – 0.73, 0.79 – 1.00	638
Genetic	11	0.05	1	0.05 – 0.94	0.00 – 0.44, 0.63 – 0.93	234
Treatment	20	0.05	1	0.05–1.00	0.00 – 0.65, 0.91–1.00	145
Psychology	50	0.07	1	0.07–1.00	0.02 – 0.24, 0.28 – 1.00	198
Imaging	32	0.11	1	0.11–1.00	0.03 – 0.54, 0.71–1.00	65
Neurochemistry	47	0.07	1	0.07–1.00	0.02 – 0.79, 0.92–1.00	50
Miscellaneous	57	0.11	1	0.11–1.00	0.09 – 1.00	38

What’s gone wrong?

The Academic Incentive Structure

We live in a publish or perish culture.

Publication number, where you publish, and citations are all used (either explicitly or implicitly) in appointment and promotion committees.

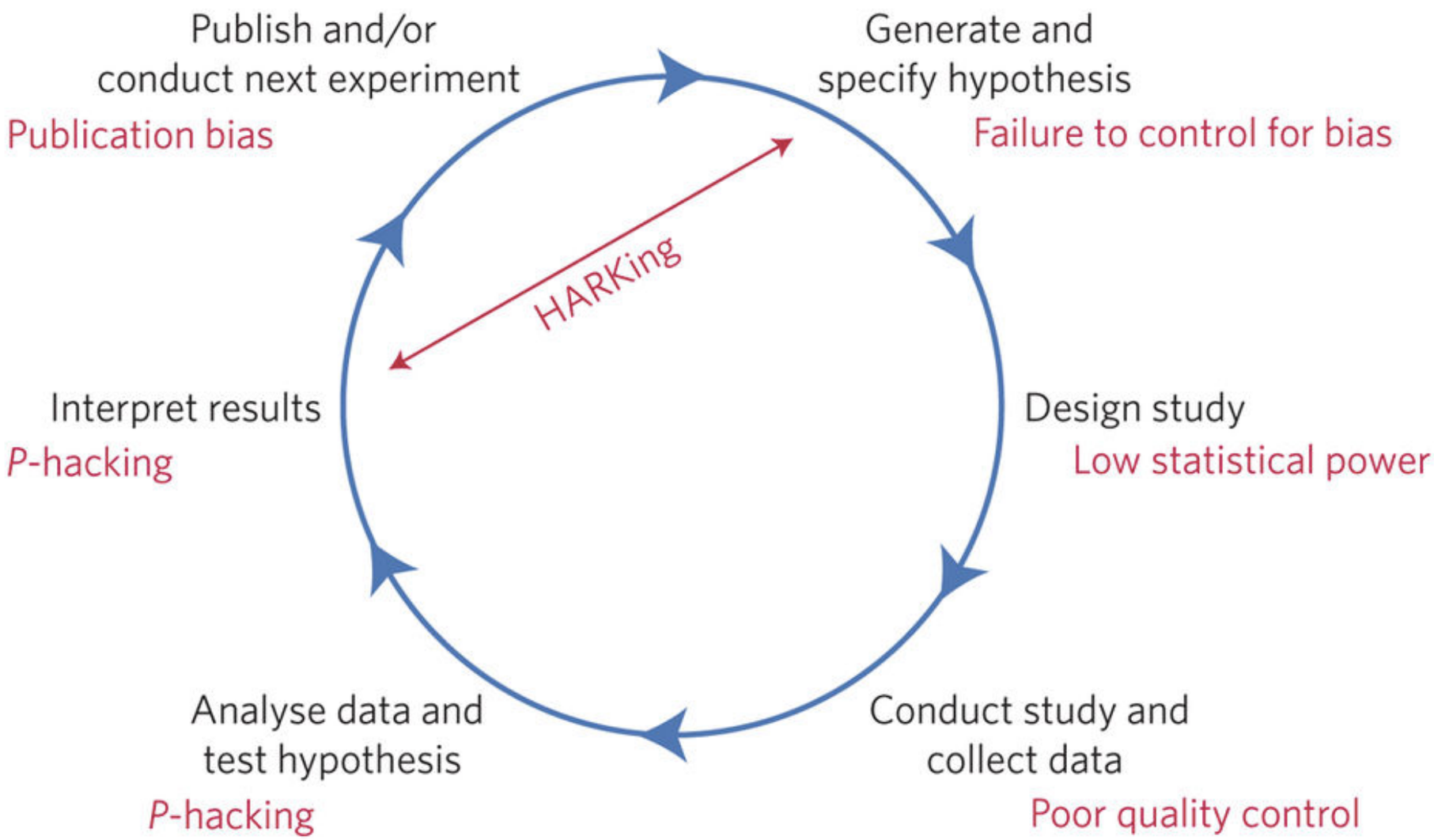
REF's definition of 3* and 4* research (although this looks like it could be changing).

Is there not just “good science” and “bad science”?

Without realising it, good scientists have been engaging in questionable research practices (QRPs) partly driven by an incentive structure that doesn’t incentivise good scientific practice…

Problems include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

From: A manifesto for reproducible science



Munafo et al. (2017), *Nature Human Behaviour*

Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKing is defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as if it were, in fact, an a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms of HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discuss several reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing's costs exceed its benefits is a complex one that ought to be addressed through research, open discussion, and debate. To help stimulate such discussion (and for those such as myself who suspect that HARKing's costs do exceed its benefits), I conclude the article with some suggestions for deterring HARKing.

ANNUAL REVIEWS

Annual Review of Psychology

Psychology's Renaissance

Leif D. Nelson,[1] Joseph Simmons,[2] and Uri Simonsohn[2]

[1]Haas School of Business, University of California, Berkeley, California 94720; email: Leif_Nelson@haas.berkeley.edu

[2]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: jsimmo@upenn.edu, urisohn@gmail.com

"the overwhelming majority of published findings are statistically significant (Fanelli 2012, Greenwald 1975, Sterling 1959). On the other hand, the overwhelming majority of published studies are underpowered and, thus, theoretically unlikely to obtain results that are statistically significant."

ROBERT TAYLOR

Rein in the four horsemen of irreproducibility

Dorothy Bishop *describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.*

More than four decades into my scientific career, I find myself an outlier among academics of similar age and seniority: I strongly identify with the movement to make the practice of science more robust. It's not that my contemporaries are unconcerned about doing science well; it's just that many of them don't seem to recognize that there are serious problems with current practices. By contrast, I think that, in two decades, we will look back on the past 60 years — particularly in biomedical science — and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses in controlled experiments. We can account for unwanted variation with statistical techniques. We appreciate the need to replicate observations.

Yet many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known). My generation and the one before us have done little to rein these in.

In 1975, psychologist Anthony Greenwald noted that science is prejudiced against null hypotheses; we even refer to sound work supporting such conclusions as 'failed experiments'. This prejudice leads to publication bias: researchers are less likely to write up studies that show no effect, and journal editors are less likely to accept them. Consequently, no one can learn from them, and researchers waste time and resources be adequately powered. Other disciplines have yet to catch up.

I stumbled on the issue of *P*-hacking before the term existed. In the 1980s, I reviewed the literature on brain lateralization (how sides of the brain take on different functions) and developmental disorders, and I noticed that, although many studies described links between handedness and dyslexia, the definition of 'atypical handedness' changed from study to study — even within the same research group. I published a sarcastic note, including a simulation to show how easy it was to find an effect if you explored the data after collecting results (D. V. M. Bishop *J. Clin. Exp. Neuropsychol.* **12**, 812–816; 1990). I subsequently noticed similar phenomena in other fields: researchers try out many analyses but report only the ones that are 'statistically significant'.

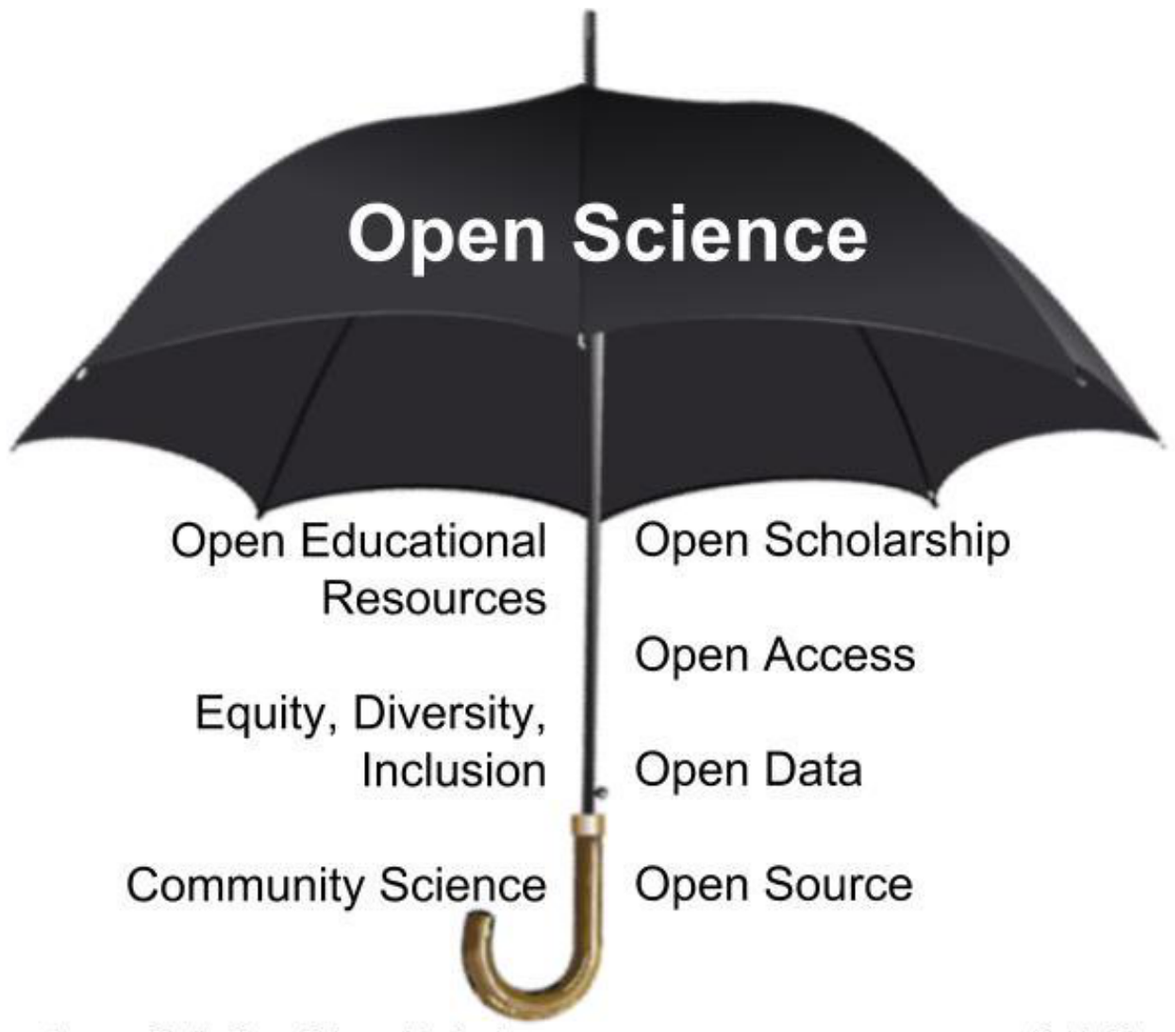
This practice, now known as *P*-hacking, was once endemic to most branches of science that rely on *P* values to test significance of results, yet few people realized how seriously it could distort findings. That started to change in 2011, with an elegant, comic paper in which the authors crafted analyses to prove that listening to the Beatles could make undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). "Undisclosed flexibility," they wrote, "allows presenting anything as significant."

The term HARKing was coined in 1998 (N. L. Kerr *Pers. Soc. Psychol. Rev.* **2**, 196–217; 1998). Like *P*-hacking, it is so widespread that researchers assume it is good practice. They look at the data, pluck out a finding that looks exciting and write a paper to tell a story around this result. Of course, researchers should be free to explore their

MANY RESEARCHERS PERSIST IN WORKING IN A WAY ALMOST GUARANTEED NOT TO DELIVER MEANINGFUL RESULTS.

https://www.nature.com/articles/d41586-019-01307-2

How can we engage in open and reproducible research?



Adapted from: https://www.meetup.com/Berlin-Open-Science-Meetup/

Robin Champieux and Danielle Robinson

Before Data Collection

- Specify your hypotheses and analysis plan.
- **Pre-register** your hypotheses and analysis plan at `osf.io`
- Consider data simulation so that you can write your analysis script before you have your real data.
- Consider submitting as a **registered report** - currently **186** journals now support this route. This involves acceptance in principle before you have even started collecting your data.

Registered Reports

https://cos.io/rr/?_ga=2.49773158.1336120275.1555407527-1361001319.1494339346

Open data, open code, open computational environment.

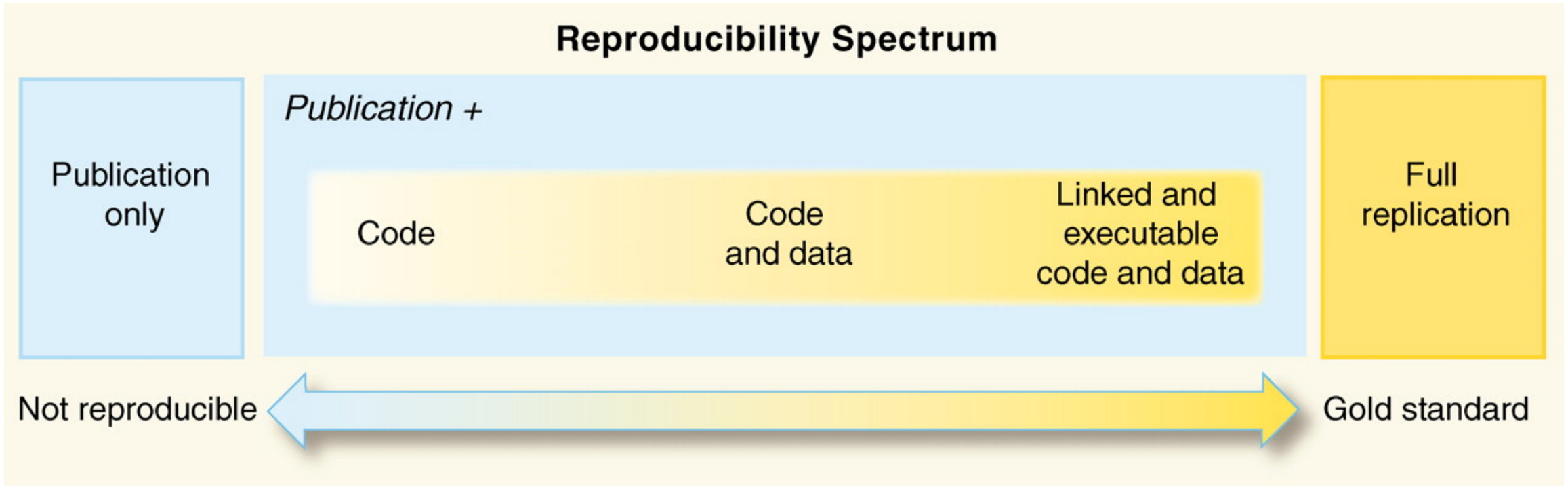
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



The UK Reproducibility Network

The power of networks

A group of researchers recently launched the UK Reproducibility Network, supported by Jisc and a range of other stakeholders, including funders and publishers.

Our aim is to bring together colleagues across the higher education and research sector, forming local networks at individual institutions to promote the adoption of initiatives intended to improve research.

> This is very much a peer-led, grassroots initiative that will allow academics to coordinate their efforts and engage with key stakeholders.

https://www.jisc.ac.uk/blog/how-do-you-deal-with-a-problem-like-reproducibility-29-nov-2018

Our Open Science Working Group

- Our Open Science Working Group founded in November by myself and Caroline.

- Connects with the UKRN.

- Advocates locally for engagement with open research practices in research and teaching.

- Check out the Network of Open Science Working groups: `https://osf.io/vgt3x/`

North West Open Science Network

- We are part of a broader network in the NW including Lancaster, Keele, MMU.

- We are also part of the UK Reproducibility Network funded/supported by UKRI, research England, MRC, NERC, ESRC, Wellcome, Universities UK, JISC, British Neuroscience Association (amongst others).

- Links to Project Tier, The Carpentries, Software Sustainability Institute, The Turing Way etc.

What we've been up to...

- ReproducibiliTea journal club every fortnight led by Jade, Richie, George, and Dan.

- Fed into University policy on Open Research/Science - paper going to senate in September.

- Recent talks at Lancaster and Keele on reproducibility.

- RUM session on Binder (reproducible computational environments).

- Several OSWG members attended/presented at SIPS.

Upcoming...

- Fortnightly ReproT.

- Reproducibility in Science afternoon - February 26th - Dorothy Bishop keynote, plus a series of 15 mins talks.

- Email invite to workshops on Declaration for Responsible and Intelligent Data Practice - academic workshop September 17th.

- Email invite to research integrity workshop commissioned by Research England - August 9th.

- But what do **you** want? Workshops to develop technical skills? Other events?