Entre Pares

OPEN SCIENCE –

OPEN RESEARCH DATA

Dr J Rogel-Salazar

IBM and Imperial College London

AGENDA

OPEN SCIENCE

- Open Movement
- Open Science v Open access
- Building blocks of OS
- Where does data fit in?
- Next steps

INTRODUCTION

ABOUT ME

- Welcome to Entre Pares
- Session on Open Science and Open Research Data

Jesús Rogel-Salazar

LET US START WITH A STORY...

EUREKA!

A DISCOVERY!

- Close your eyes and keep calm…
- Imagine that a dear colleague of yours has sent you the following email:

I've found something amazing. I don't have time to tell you exactly what it is, or how I found it, but here's proof that I discovered it:

smaismrmilmepoetaleumibunenugttauiras

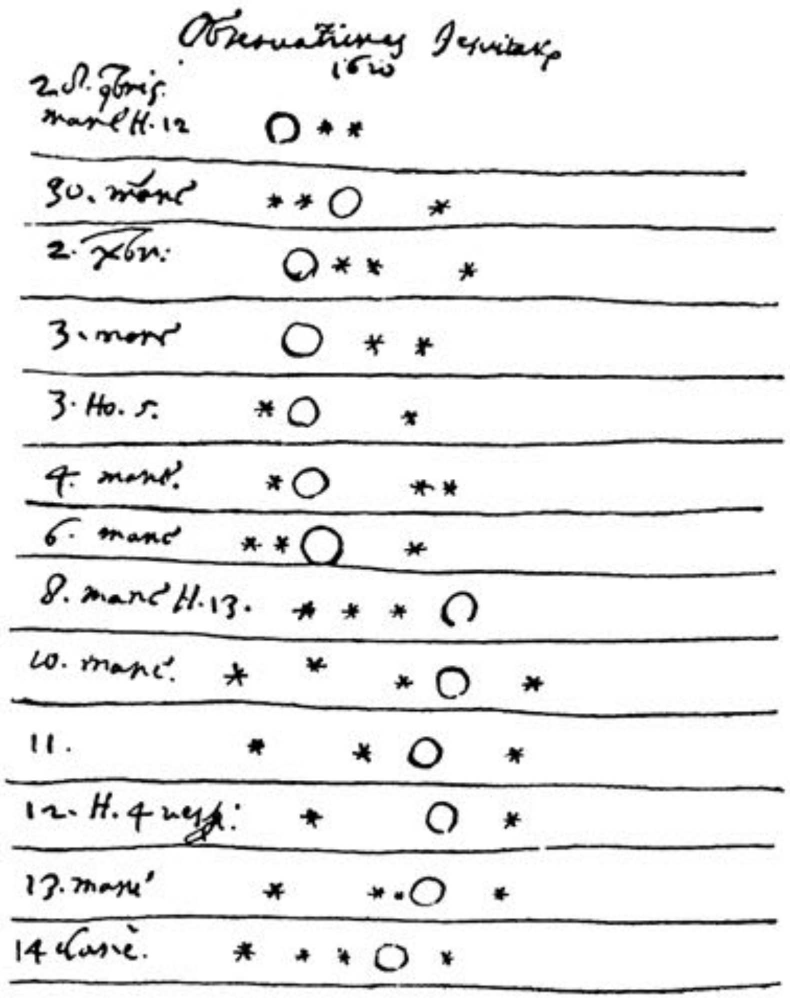
Eureka!

Yours truly,

GALILEO

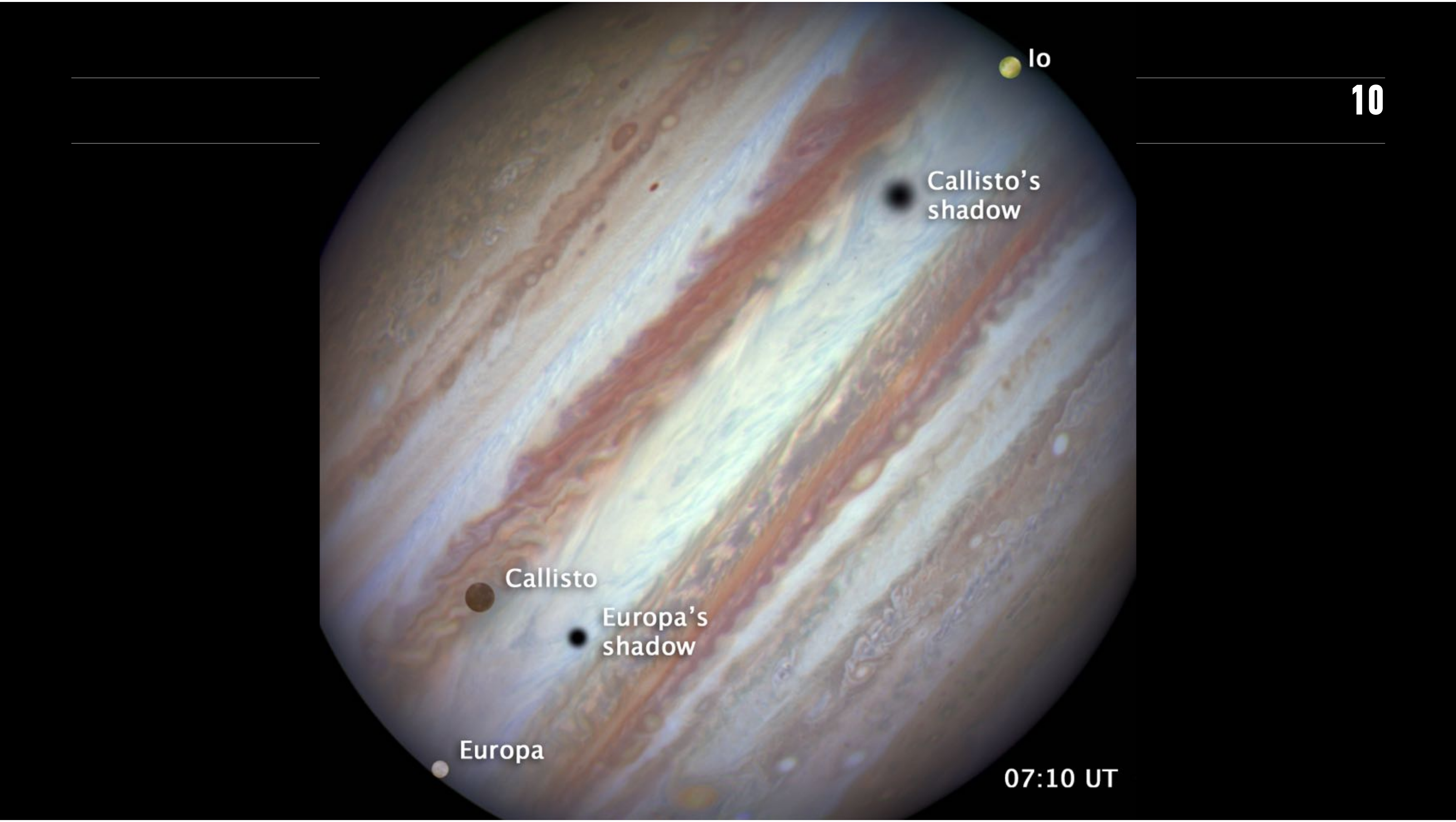
- He did not shout Eureka after his discovery. That was Archimedes, but…
- On the 25th of July in 1610, he discovered that Saturn was apparently situated between two smaller companions that always moved together
- Wanting to establish his priority of discovery, he sent to Kepler (and others) the following anagram, which he informed them was a coded description of his latest discovery:

smaismrmilmepoetaleumibunenugttauiras


1610
2 S. Gbris. mane H. 12
30. mane
2. xbris
3. mane
3. Ho. s.
4. mane
6. mane
8. mane H. 13.
10. mane
11.
12. H. 4 vesp:
13. mane
14 Cane.

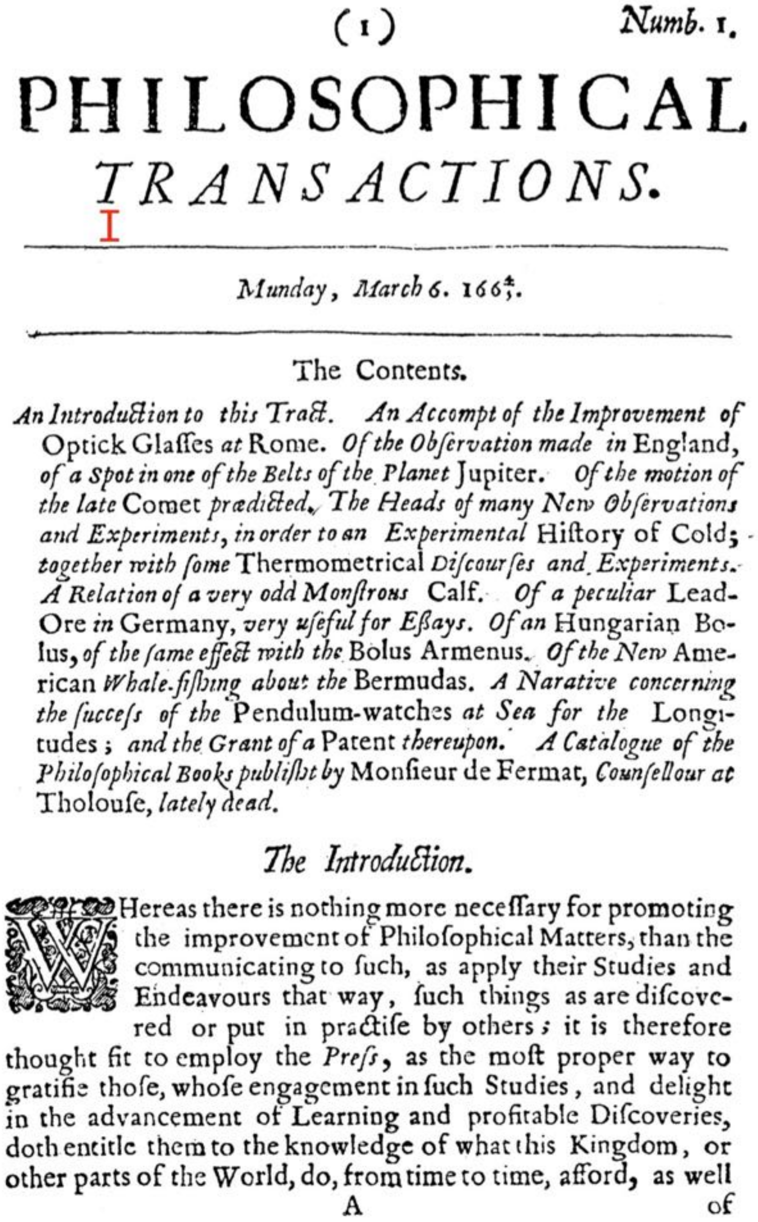
ALTISSIMUM PLANETAM TERGEMINUM OBSERVAVI

I HAVE OBSERVED THE HIGHEST OF THE PLANETS [JUPITER] THREE-FORMED


Io
Callisto's
shadow
Callisto
Europa's
shadow
Europa
07:10 UT

COMMON PRACTICE... BACK THEN?

- Sadly, this kind of scientific communication was common at the time
- Newton, Huygens, Hooke, and Leonardo all used similar devices to hide their discoveries and methods from each other
- In 1665, the Philosophical Transactions (one of the earliest scientific journals) was founded by Henry Oldenburg.



(1) *Numb.* 1.

PHILOSOPHICAL *TRANSACTIONS.*

Munday, March 6. 166$\frac{4}{5}$.

The Contents.

An Introduction to this Tract. An Accompt of the Improvement of Optick Glaſſes *at* Rome. *Of the Obſervation made in* England, *of a Spot in one of the Belts of the Planet* Jupiter. *Of the motion of the late* Comet *prædicted. The Heads of many New Obſervations and Experiments, in order to an Experimental* Hiſtory of Cold; *together with ſome* Thermometrical *Diſcourſes and Experiments. A Relation of a very odd Monſtrous* Calf. *Of a peculiar* Lead-Ore *in* Germany, *very uſeful for Eßays. Of an* Hungarian Bolus, *of the ſame effect with the* Bolus Armenus. *Of the New* American *Whale-fiſhing about the* Bermudas. *A Narative concerning the ſucceſs of the* Pendulum-watches *at Sea for the* Longitudes; *and the Grant of a* Patent *thereupon. A Catalogue of the Philoſophical Books publiſht by* Monſieur de Fermat, *Counſellour at* Tholouſe, *lately dead.*

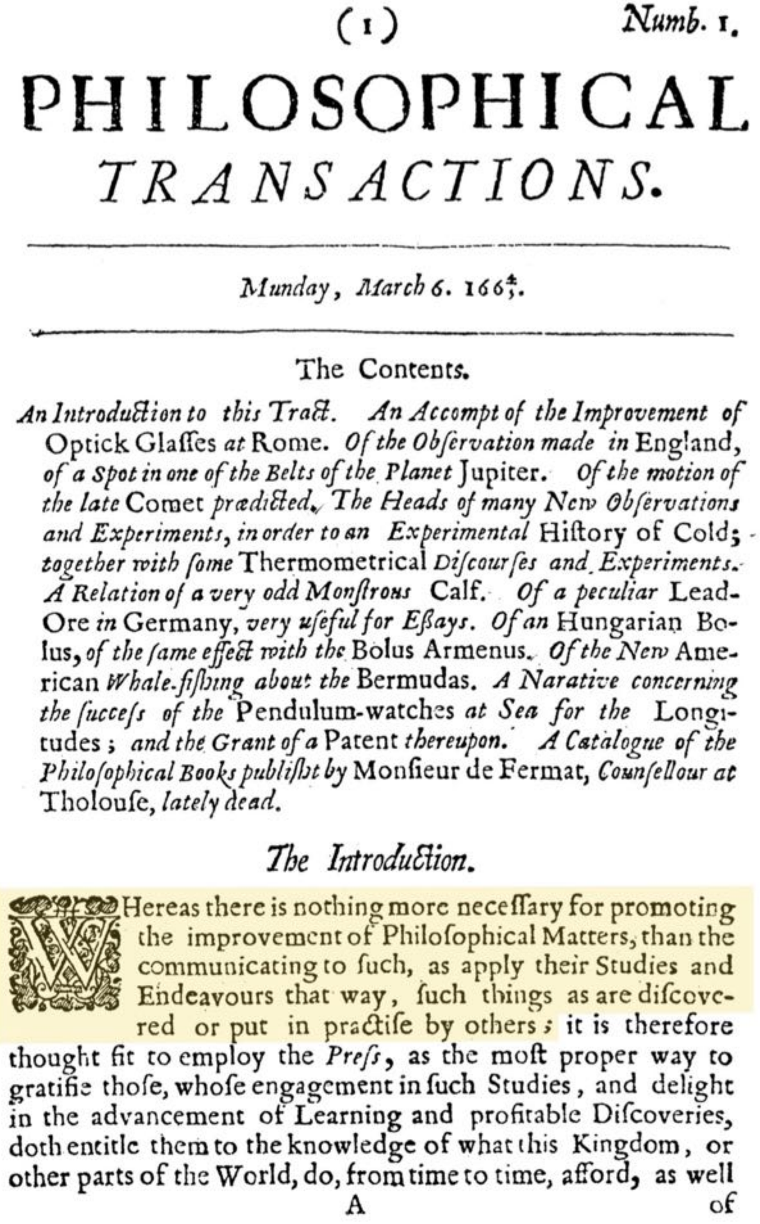
The Introduction.

WHereas there is nothing more neceſſary for promoting the improvement of Philoſophical Matters, than the communicating to ſuch, as apply their Studies and Endeavours that way, ſuch things as are diſcovered or put in practiſe by others; it is therefore thought fit to employ the *Preſs*, as the moſt proper way to gratifie thoſe, whoſe engagement in ſuch Studies, and delight in the advancement of Learning and profitable Diſcoveries, doth entitle them to the knowledge of what this Kingdom, or other parts of the World, do, from time to time, afford, as well

A of

COMMON PRACTICE… BACK THEN?

- Sadly, this kind of scientific communication was common at the time
- Newton, Huygens, Hooke, and Leonardo all used similar devices to hide their discoveries and methods from each other
- In 1665, the Philosophical Transactions (one of the earliest scientific journals) was founded by Henry Oldenburg.



(1) Numb. 1.

PHILOSOPHICAL
TRANSACTIONS.

Munday, March 6. 166$\frac{4}{5}$.

The Contents.

An Introduction to this Tract. An Accompt of the Improvement of Optick Glaſſes *at* Rome. *Of the Obſervation made in* England, *of a Spot in one of the Belts of the Planet* Jupiter. *Of the motion of the late* Comet *prædicted. The Heads of many New Obſervations and Experiments, in order to an Experimental* Hiſtory of Cold; *together with ſome* Thermometrical *Diſcourſes and Experiments. A Relation of a very odd Monſtrous* Calf. *Of a peculiar* Lead-Ore *in* Germany, *very uſeful for Eßays. Of an* Hungarian Bolus, *of the ſame effect with the* Bolus Armenus. *Of the New* American *Whale-fiſhing about the* Bermudas. *A Narative concerning the ſucceſs of the* Pendulum-watches *at Sea for the* Longitudes; *and the Grant of a* Patent *thereupon. A Catalogue of the Philoſophical Books publiſht by* Monſieur de Fermat, *Counſellour at* Tholouſe, *lately dead.*

The Introduction.

WHereas there is nothing more neceſſary for promoting the improvement of Philoſophical Matters, than the communicating to ſuch, as apply their Studies and Endeavours that way, ſuch things as are diſcovered or put in practiſe by others; it is therefore thought fit to employ the *Preſs*, as the moſt proper way to gratifie thoſe, whoſe engagement in ſuch Studies, and delight in the advancement of Learning and profitable Diſcoveries, doth entitle them to the knowledge of what this Kingdom, or other parts of the World, do, from time to time, afford, as well

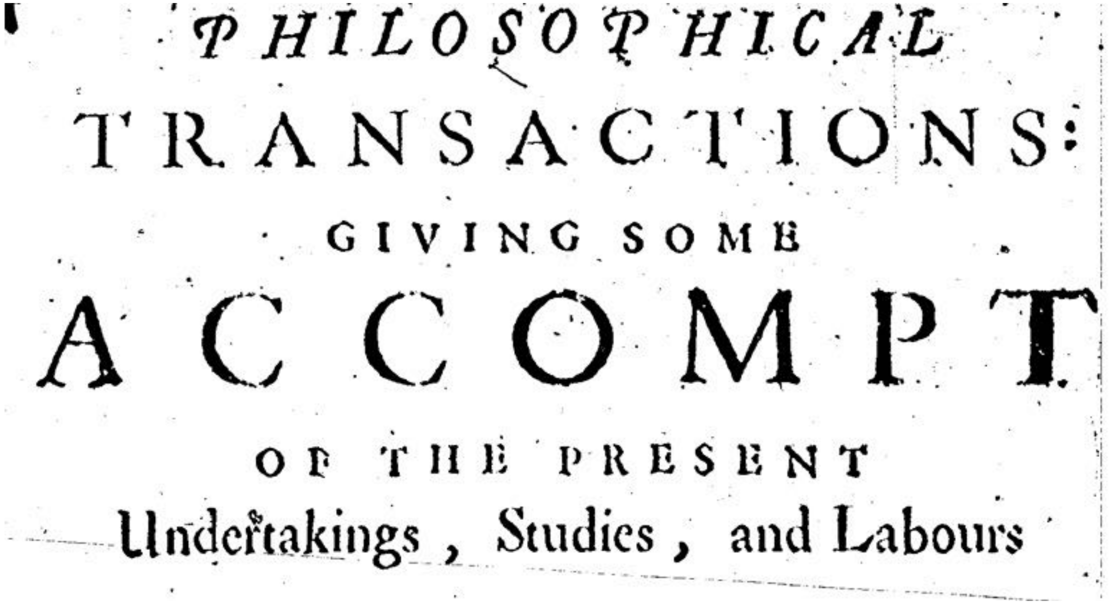
A of

WORK, FINISH, PUBLISH

- In Michael Faraday’s obituary by William Crookes we hear the advice that the famous researcher gave to his students:

  “The secret is comprised in three words — Work, Finish, Publish.”

- To this date this rings a bell, except for the few occasions when some young students firmly omit the second verb.



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours

DOING SCIENCE...

SOME CHANGES SINCE 1665

HOWEVER…

- Nowadays, there are thousands of papers and reports that, for example, tell us about computations that cannot be reproduced without access to secret software
- The secret sauce of that software is hidden from other researchers
- Reproducibility is hindered, and depending on the capacity of the owner of that secret sauce, you may be at odds with licensing infringements and the like

SCIENCE FOR ALL…

- Science has been open since 1665
- We just need to remind ourselves, our colleagues and institutions that this is the case!

INTRODUCTION

OPEN MOVEMENT

OPEN

WHAT DOES IT MEAN?

IT DEPENDS

OPEN ACCESS TO PUBLICATIONS

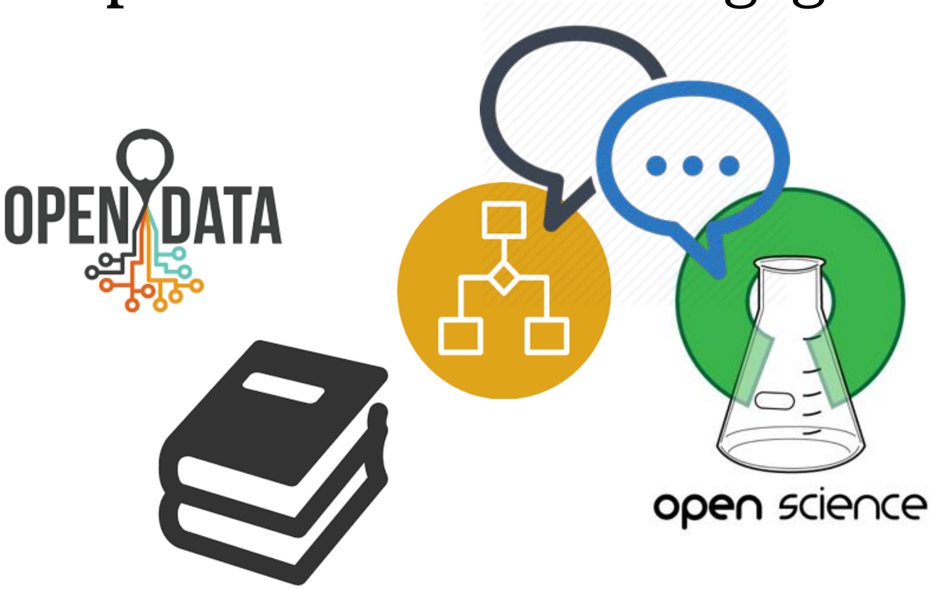
- Free, immediate, online access to the results of research
- Make sure anyone can access your papers
  - Well established concept
- **Gold route**: paying Article Publishing Charges to ensure publisher makes copy open
- **Green route**: self-archiving Open Access copy in repository
- Find out what your publisher allows on SHERPA RoMEO
  - www.sherpa.ac.uk/romeo

HOWEVER...

OPEN ACCESS DOES NOT EQUAL OPEN SCIENCE

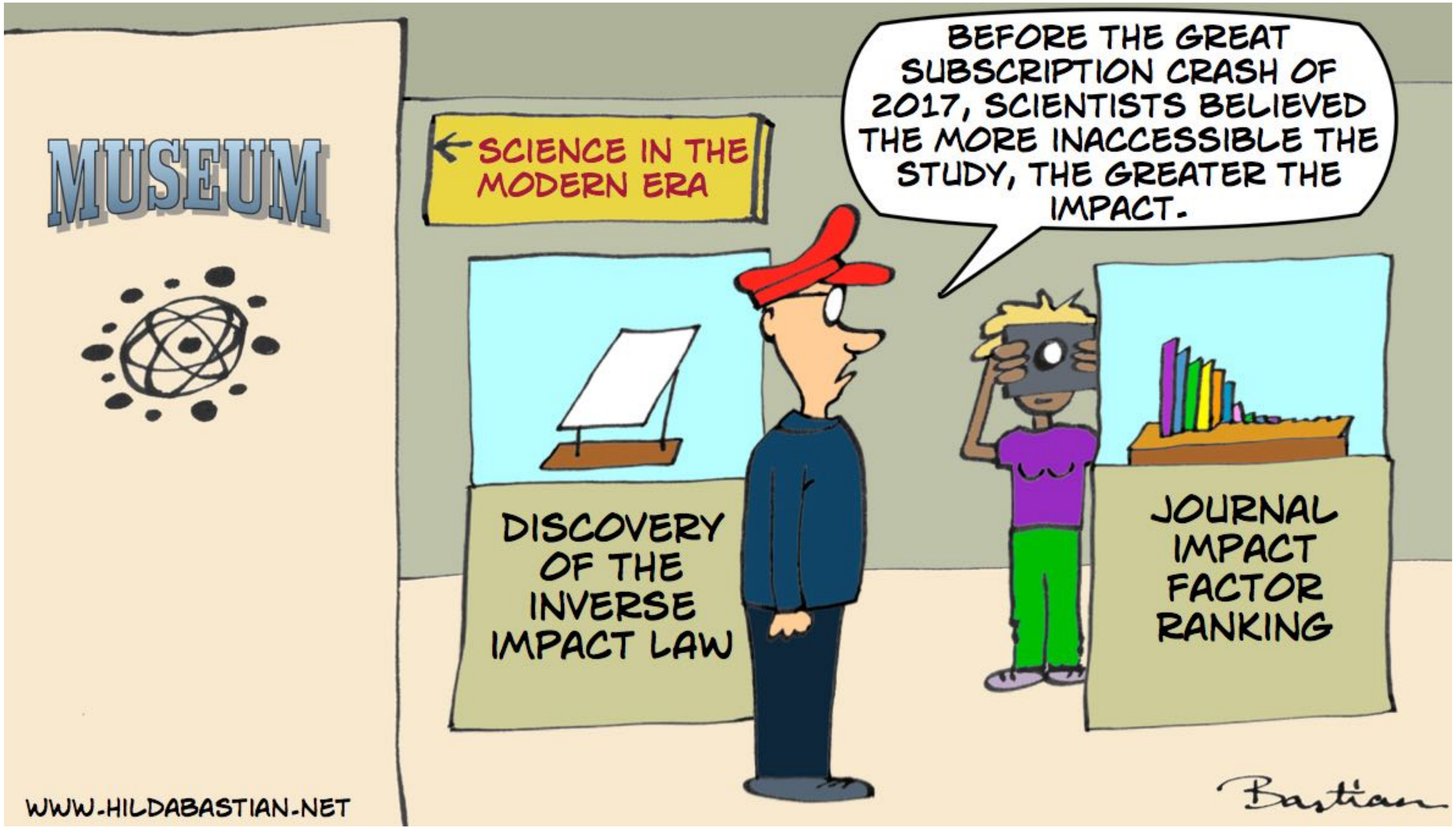
OPEN SCIENCE VS OPEN ACESS

- Transparent processing and methods
- Open data
- Community science
- Open discussion and engagement



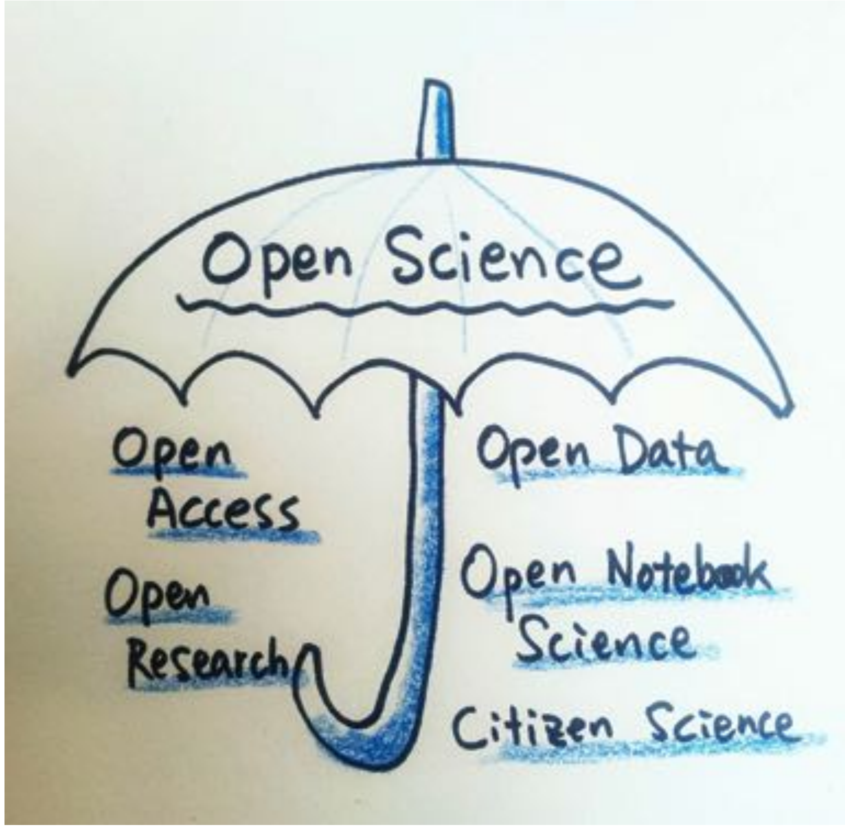
- Gold or green OA
- Open Access literature is “digital, online, free of charge, and free of most copyright and licensing restrictions”
- Free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles…

MEASURING IMPACT


MUSEUM
← SCIENCE IN THE MODERN ERA
BEFORE THE GREAT SUBSCRIPTION CRASH OF 2017, SCIENTISTS BELIEVED THE MORE INACCESSIBLE THE STUDY, THE GREATER THE IMPACT.
DISCOVERY OF THE INVERSE IMPACT LAW
JOURNAL IMPACT FACTOR RANKING
WWW.HILDABASTIAN.NET
Bastian

OPEN SCIENCE

- Offers researchers tools and workflows for
  - **transparency**
  - **reproducibility**
  - **dissemination** and
  - transfer of new knowledge
- OS conducts enables **collaboration and contribution:**
  - **research data, lab notes and other research processes** are freely available,
  - with terms that allow reuse, redistribution and reproduction of the research



OPEN METHODS

- Computer code, methods and tools
  - Shared to allow other to reproduce work
- Facilitate collaboration by
  - Sharing workflows and documentation
  - using web based tools
    - Open netbook science – Freely available URL to a lab’s notebook, indexed on common search engines

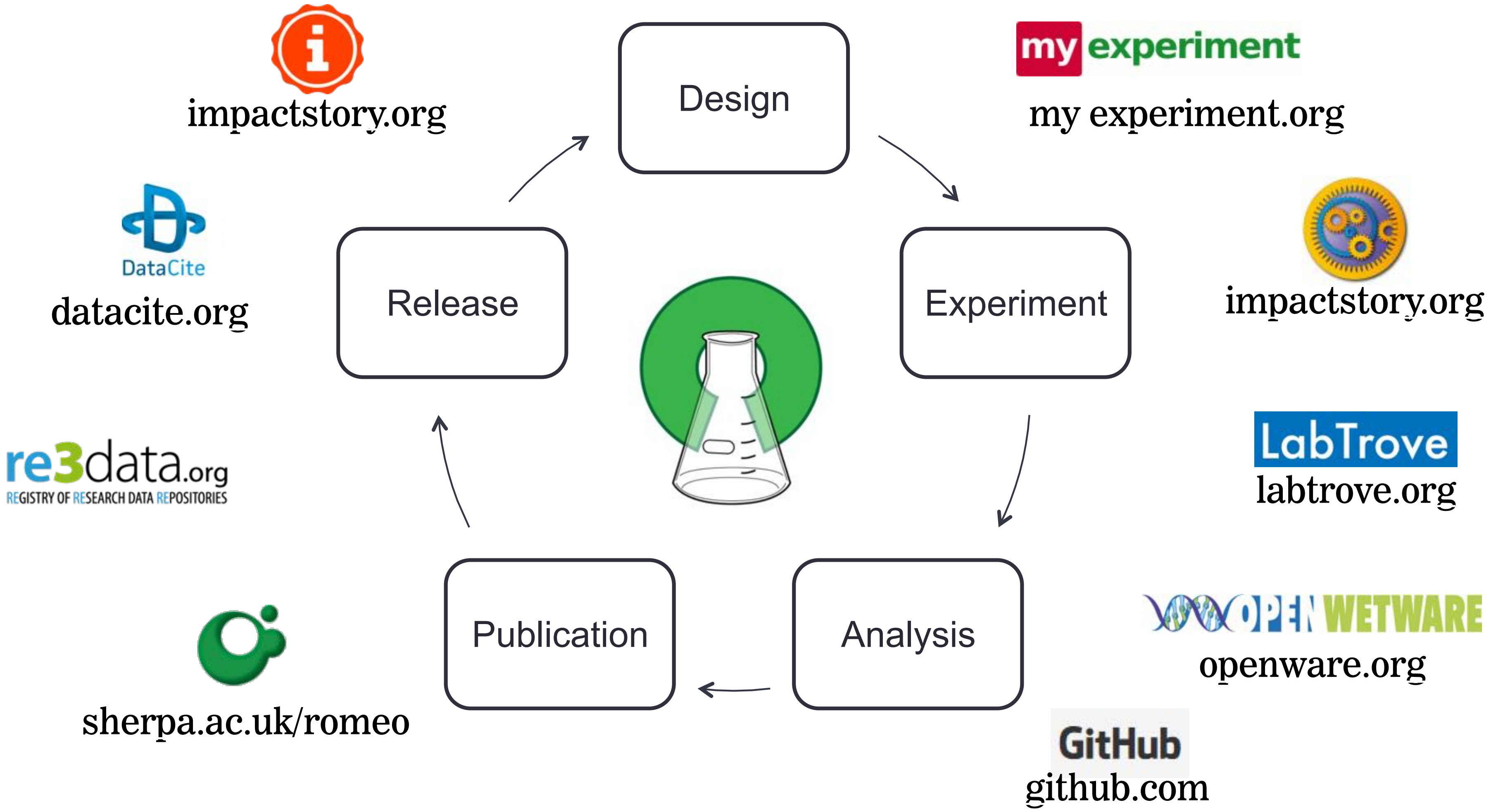
OPEN DATA

- Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike
- Enable interoperability
  - ability of diverse systems and organisations to work together
  - In this case, it is the ability to interoperate - or intermix - different datasets.
- Include metadata!

OPEN SCIENCE COMPONENTS

- Open Methodology (Methods, processes, relevant documents)
- Open Source (Soft- & Hardware)
- Open Data (data free to re-use)
- Open Access to scholarly outputs
- Open Peer Review (transparency in evaluation and quality criteria)
- Open Educational Resources (MOOCs, OERs)

OPEN AT EVERY STEP OF THE RESEARCH



Design

Experiment

Analysis

Publication

Release

impactstory.org

my experiment.org

datacite.org

impactstory.org

labtrove.org

openware.org

github.com

sherpa.ac.uk/romeo

OPEN SCIENCE LANDSCAPE

- Collaboration
- New forms of peer review
- Open infrastructure
- Research data management
- Open access publishing
- Massive Open Online Courses (MOOCs)

- Alternative Metrics
- Policy
- Open data
- Open science
- Copyright & Licensing
- Open education
- Advocacy & training

OPEN DATA

open science

NEVER PART OF GOOD RESEARCH PRACTICES ANYWAY…

"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

Original image via doi:10.1038/461145a. "Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science." - *Nature* 461, 145

WHAT’S IN IT FOR ME?

BENEFITS OF OPEN SCIENCE

BENEFITS OF OPEN SCIENCE

- Increase the visibility of your research and its impact
- Transparency and reproducibility
- Foster new collaborations and research dialogue
- Access to relevant literature
  - not behind pay walls
  - in a faster manner
- Increase the efficiency of research
- Open to new ideas, methods and processes

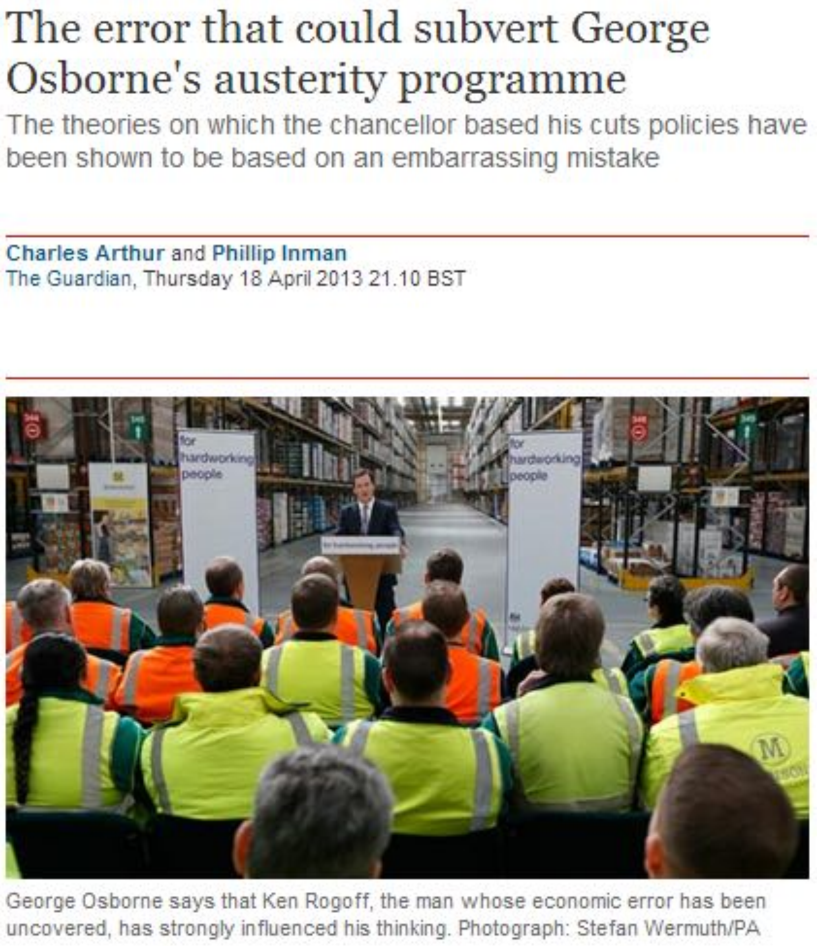
BENEFITS OF OPEN SCIENCE

- Validation of results
- Tackle academic fraud
  - Reproducibility and accountability
- Acceleration of research process
- More scientific breakthroughs
- Increased use and economic benefit

VALIDATION OF RESULTS

“It was a mistake in a spreadsheet that could have been easily overlooked: a few rows left out of an equation to average the values in a column.

The spreadsheet was used to draw the conclusion of an influential 2010 economics paper: that public debt of more than 90% of GDP slows down growth. This conclusion was later cited by the International Monetary Fund and the UK Treasury to justify programmes of austerity that have arguably led to riots, poverty and lost jobs.”



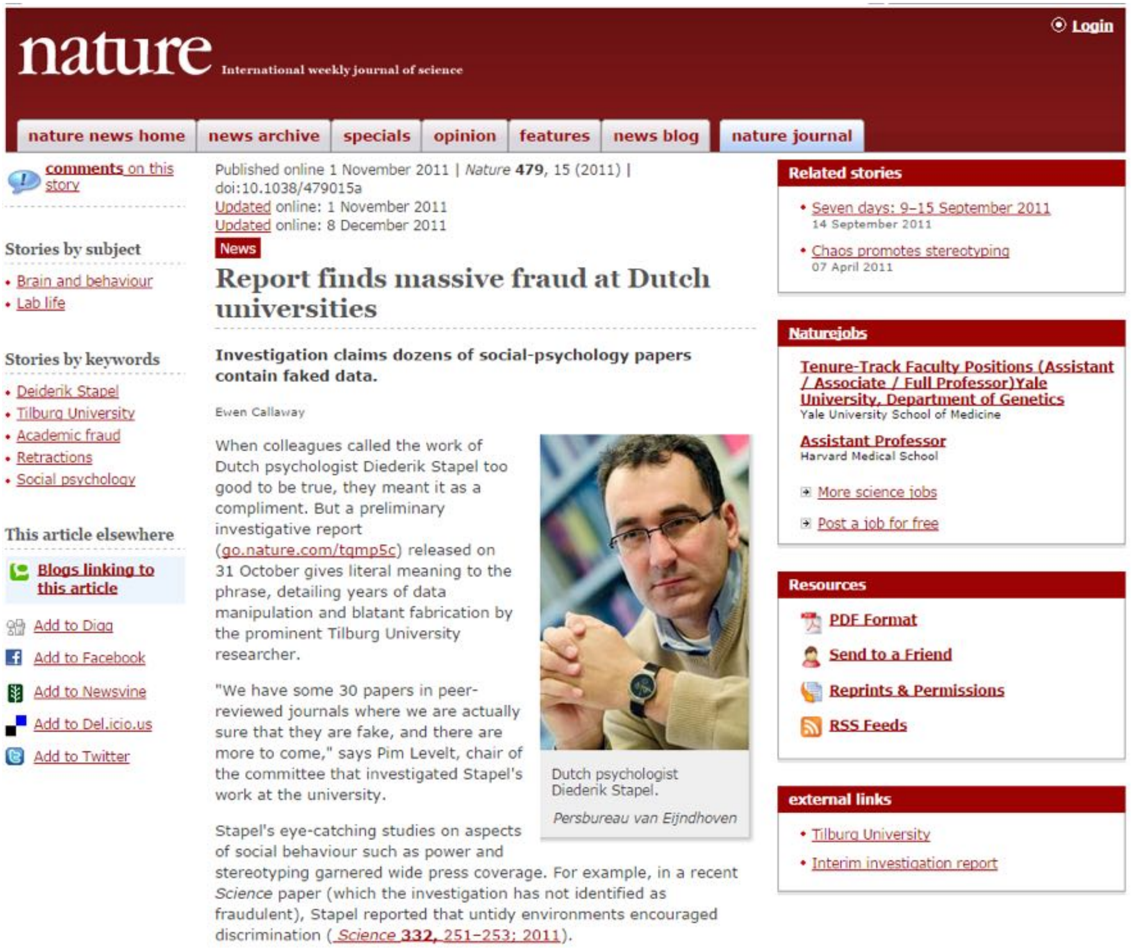
The error that could subvert George Osborne's austerity programme

The theories on which the chancellor based his cuts policies have been shown to be based on an embarrassing mistake

Charles Arthur and **Phillip Inman**
The Guardian, Thursday 18 April 2013 21.10 BST

George Osborne says that Ken Rogoff, the man whose economic error has been uncovered, has strongly influenced his thinking. Photograph: Stefan Wermuth/PA

CUT DOWN ACADEMIC FRAUD



nature International weekly journal of science

⦿ Login

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject

- Brain and behaviour
- Lab life

Stories by keywords

- Deiderik Stapel
- Tilburg University
- Academic fraud
- Retractions
- Social psychology

This article elsewhere

Blogs linking to this article

Add to Digg

Add to Facebook

Add to Newsvine

Add to Del.icio.us

Add to Twitter

Published online 1 November 2011 | *Nature* **479**, 15 (2011) | doi:10.1038/479015a

Updated online: 1 November 2011

Updated online: 8 December 2011

News

Report finds massive fraud at Dutch universities

Investigation claims dozens of social-psychology papers contain faked data.

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tgmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

"We have some 30 papers in peer-reviewed journals where we are actually sure that they are fake, and there are more to come," says Pim Levelt, chair of the committee that investigated Stapel's work at the university.

Dutch psychologist Diederik Stapel.

Persbureau van Eijndhoven

Stapel's eye-catching studies on aspects of social behaviour such as power and stereotyping garnered wide press coverage. For example, in a recent *Science* paper (which the investigation has not identified as fraudulent), Stapel reported that untidy environments encouraged discrimination (*Science* **332**, 251–253; 2011).

Related stories

- Seven days: 9–15 September 2011
  14 September 2011
- Chaos promotes stereotyping
  07 April 2011

Naturejobs

Tenure-Track Faculty Positions (Assistant / Associate / Full Professor)Yale University, Department of Genetics
Yale University School of Medicine

Assistant Professor
Harvard Medical School

More science jobs

Post a job for free

Resources

PDF Format

Send to a Friend

Reprints & Permissions

RSS Feeds

external links

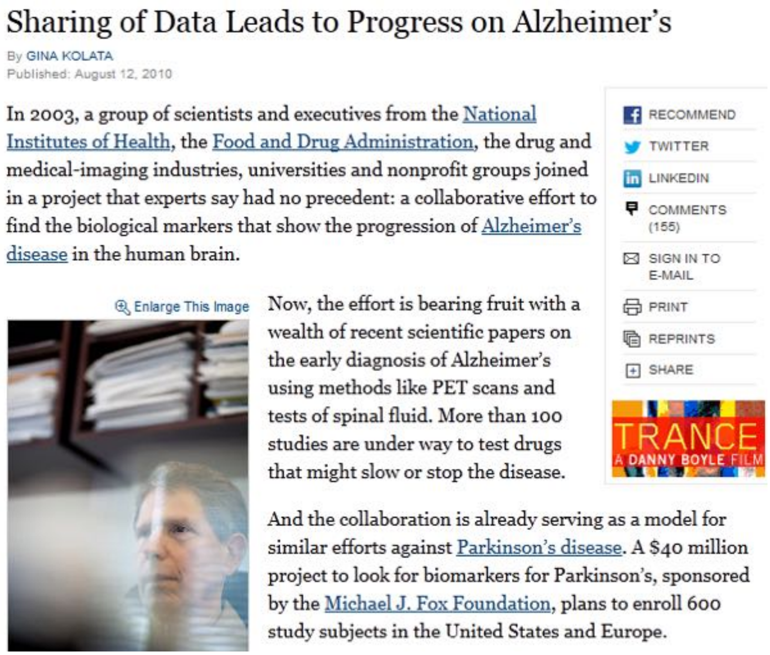
- Tilburg University
- Interim investigation report

www.nature.com/news/2011/111101/full/479015a.html

MORE RESEARCH BREAKTHROUGHS

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania



Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

INCREASED USE AND ECONOMIC BENEFIT

NASA LANDSAT SATELLITE

Up to 2008

- Sold through the US Geological Survey for $600 USD per scene
- Sales of 19k scenes/yr
- Annual revenue of $11.4M USD

Since 2009

- Freely available over the internet
- Google Earth uses the images
- Transmission of 2,100,000 scenes per year
- Estimated to have created value for the environmental management industry of $935 million, with direct benefit of more than $100 million per year to the US economy
- Has stimulated the development of applications from a large number of companies worldwide

OPEN SCIENCE

BUILDING BLOCKS

OPEN ACCESS

- Accessible
  - make it available in accessible formats (XML)
- Findable
  - put it in open and sustainable infrastructure
- Reusable
  - attach clear permissions statements/ licences

OPEN DATA

- Data must be:
  - Open by default (G8 Open Data Charter)
  - Usable by all
  - Available
  - Findable
  - Interpretable
  - Citable
  - Curated/preserved

SKILLS AND TRAINING

The transition to an open science paradigm where research data plays a significant role requires training and education for researchers and for data managers who support open science

- Need to embed training in post graduate education
- Invest in the development of the data professional
- Training provision as and when needed (importance of train the trainer)
- Training and support for new tools and methods

EVEN BY PLAYING… DATAPOLIS

Datopolis is a board game by Ellen Broad and Jeni Tennison from the Open Data Institute, and as you might expect, it promotes the use of open data

A game about building things — services, websites, devices, apps and research — using closed and open data.

You can also download the components from GitHub and print them yourself, because it’s all openly licensed, fittingly.

https://github.com/opendataboardgame/game

INFRASTRUCTURE

Open collaborative and interoperable infrastructure for access to, exploitation, reuse, and the preservation of research outputs is a key enabler of Open Science.

- Continue building such infrastructure
- Foster stakeholders's trust in this infrastructure - national and international
- Support to develop and adopt global standards for interoperability.

ADVOCACY

Advocacy for Open Science, and associated societal and economic benefits, can act as an enabler as it will engender buy-in from policy makers and from researchers themselves

- Advocate for roadmaps and policies that promote open science at institutional and national level
- Advocate for changes in practice e.g. data citation, use of cc licences
- Promote your Open Science project
- Engage new audiences

INCENTIVES

- Need to discuss current system of incentives and assessment
- Supplement journal based metrics with other measures
- Consider value and impact of ALL research outputs (data, software…)
- Align assessment with institutional values
- Only a change of system of incentives will truly change practice and culture

POLICY

- Legal clarity
- Interoperability
- Ensure researchers have right to secondary publication
- Standard open access licences
- CC-by and CC0/PD

REPRODUCIBLE COMPUTATIONAL SCIENCE

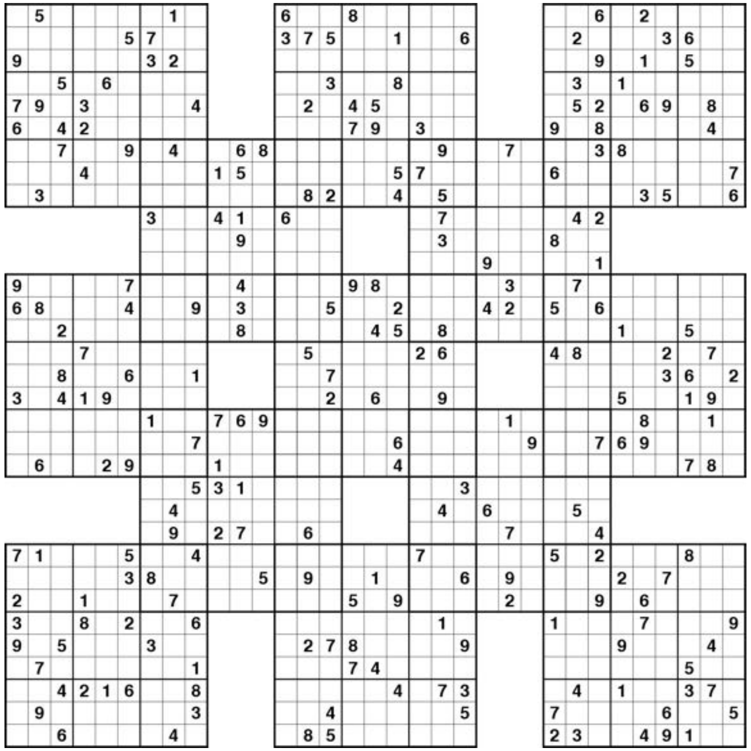
REPRODUCIBILITY IN COMPUTATIONAL SCIENCE

- Simple models and small datasets, calculations are reproducible in principle and in practice
- As simulations become more complex and datasets become larger, calculations that are reproducible in principle are no longer reproducible in practice without access to the code, data, and meta-data
- Reproducibility now requires **public access to code, data, and meta-data**

NUMERICAL EXPERIMENTS

- All source code needed to reproduce the calculation
- All input data used to perform the calculation
- All meta-data required to allow other codes to use the input data

These are equivalent to the methodology section of an experimental paper. This standard requires Open Source, Open Data, and Open Meta-data



REPRODUCIBLE RESEARCH - STANDARDS

- Release media components (text, figures) under CC-BY
- Release code components under MIT license or similar
- Attribution license on selection and arrangement of data
- Release data under CC0.

OPEN SCIENCE

WHY AREN'T ALL SCIENTIFIC PROGRAMMES OPEN SOURCE?

AT LEAST ONCE A YEAR, RESEARCHERS NEED TIME TO DO THIS:

Generate Personal "Citation Report"

1. Go to ISI Web of Science (or similar)
2. In the search box type in:
   - Last name and first initial (no commas). Use all variations, i.e. Surname N OR Surname NM
   - Make sure that Author appears in the small box on the right
3. Select the desired period of time ("All years" may be the default) and start the search
4. You can refine the results helping you limit the list to only your citations
5. Generate the Citation Report. Make sure you remove any citations that are not yours
6. Copy-paste the results in to your own report, don't forget to include citation metrics (h-index, citations, etc.)
7. Do a pirouette
8. You are done! (Don't forget to Log Out and see you next time)

COME TO THE WORKSHOP ON TUE AT 15.30, ROOM 12

RECOGNITION & ATTRIBUTION

- Researchers are measured by publishing
  - Paper count
  - Citation count
  - h-index
- Time spent on open science projects reduces publication rates
  - Scientific software tools are often not cited
  - Even if they were cited, how would that citation get tied to a researcher?
  - How can a scientist show her institution the value of her project?

ATTRIBUTION METRICS

- Should take into account:
  - Effort to maintain a useful resource
  - Importance to the scientific community
  - Externalities beyond the scientific community
- Until recently, there was no way to measure open products of research (outside of traditional publications) with a simple metric that can be used by institutions.
- This is starting to change
  - ImpactStory, DOI lookups

SUSTAINABILITY - GOOD CODING PRACTICES ARE RARE AMONG RESEARCHERS (SADLY)

Because they aren’t forced into good practices, scientists often create code that is impossible to maintain effectively. This does not lead to sustainable open science.

- Source version control systems (cvs, svn, git, Hg)
- Agile (or any other) development models
- Design patterns
- Object-oriented languages
- Public source repositories (SourceForge, github)
- Differences among open source licenses
- Unit testing
- Bug & issue tracking
- Designing for usability and usability testing
- UI design
- Error handling
- Introductory user documentation

SUSTAINABILITY – WHAT ABOUT PROFESSIONAL DEVELOPERS

Possible but

- must take into account the cost and training
- may know little about the domain

It may be significantly faster to train a computational physicist in good coding practices than it is to train even an accomplished programmer in the various disciplines we use.

SUSTAINABILITY - RESOURCES

- Tools are rarely funded
  - There is little room for projects which enrich the overall scientific enterprise, but don’t constitute novel research themselves
- Funding agencies should require delivery of primary research products:
  - code in a public repository
  - data in a public repository
  - make depositions a part of the reporting structure for funded grants

PRACTICAL TOOLS FOR OPEN SCIENCE

OPTIONS

- For papers, the standard in this community is arXiv
  - See self-archiving policies
  - Open access journals

- Social networking:
  - Mendeley
  - academia.edu
  - ResearchGate
  - Google Scholar

OPTIONS

- Sharing other output:
  - FigShare
  - SlideShare
  - GitHub

- Q&A sites:
  - Quora
  - StackOverflow
    - *Whatever*Overflow (Maths, Physics, Programmers, Baking & what-not)

GITHUB

- Became the de facto open source site
- Quick to put things online: code, notebooks, figures, data, blog, website
- Ease of collaboration: other people can build on your work with low effort
- Stars, followers, forks: Good work will attract attention.
- Citable code (DOI):Notebooks render nicely
- Reflects scientific thinking and workflow a lot better than social networks designed for researchers

SHARING & COLLABORATING

- Literate programming: Notebooks
  - Mix text, mathematical formulas, code, figures and data in the same context

- It is outstanding for explaining ideas to others, it can serve as a computational appendix.

- It is not so good for development.

- Code: how actual development is done. It helps others to continue your work.

BLOG-AWARE WEBSITE

- Static website: secure and fast
  - GitHub: username.github.io
  - WordPress
  - Jekyll
  - Pelican

CONCLUSION

CLOSING REMARKS

- Science has evolved a lot since the 1600s
- The open movement is here to stay
- But… Open Access is not Open Science
- Open Science involves:
    - Transparent processing and methods
    - Open data
    - Community science
    - Open discussion and engagement
- There are a number of tools out there to get you started

ENTRE PARES

Q&A