Open data: nice people can’t share!

Dominique Roche
Université de Neuchâtel, Switzerland
@dom_roche

Université de Neuchâtel

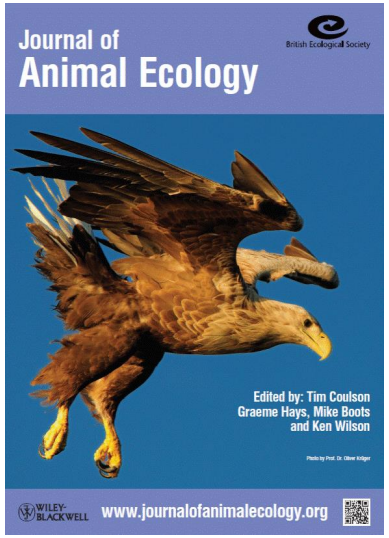
Carleton UNIVERSITY
Canada’s Capital University

COS

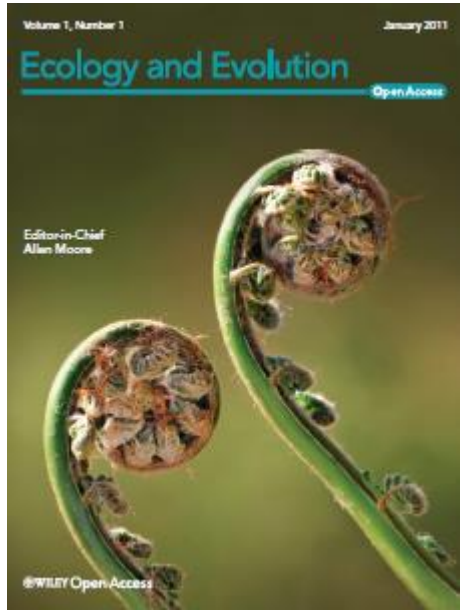
CENTER FOR OPEN SCIENCE

figshare

credit for **all** your research



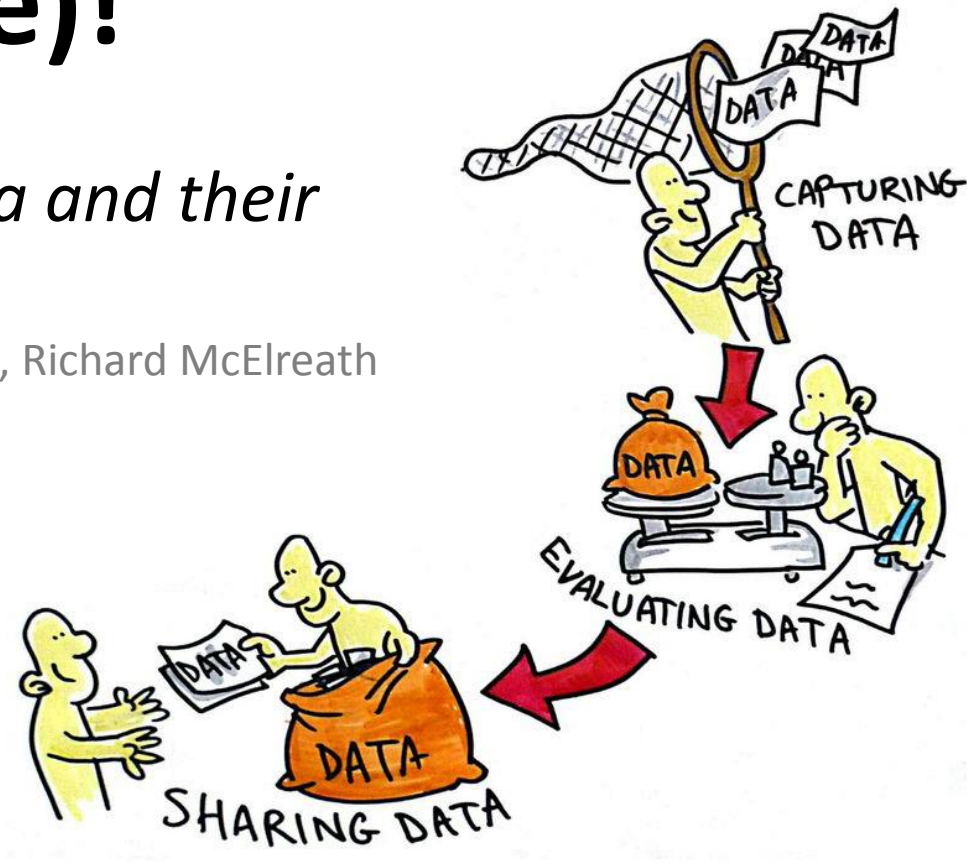
Publications are not the only useful research output!



DATA MATTER (more)!

"The paper is just an advertisement. The data and their analysis are the scientific product."

Victoria Stodden, Richard McElreath

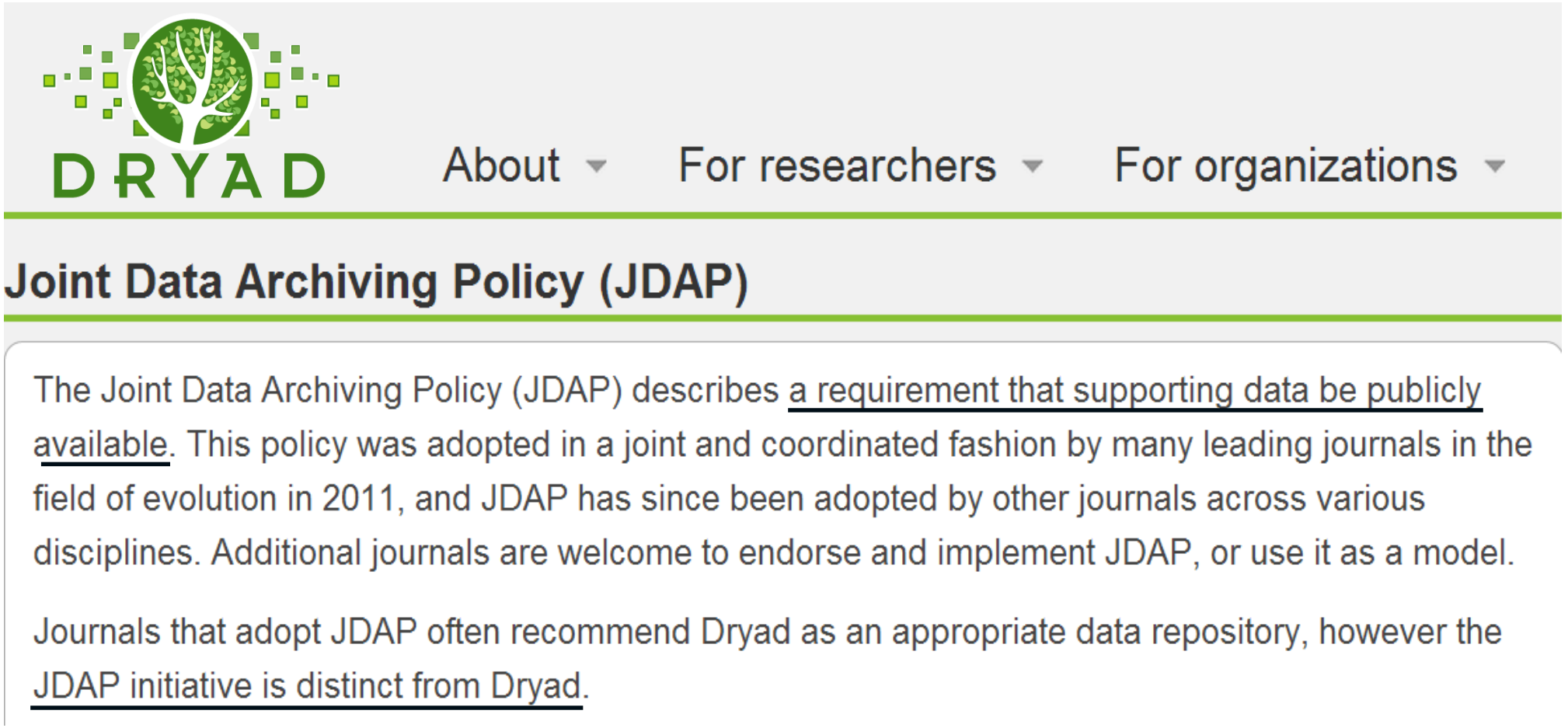


Do scientists have an obligation to make their data freely available?

Big push in the biological sciences for

OPEN DATA

Joint Data Archiving Policy (JDAP)



http://datadryad.org/pages/jdap

Journals that require open data

Examples:

- The American Naturalist
- Biological Journal of the Linnean Society
- Biology Letters
- BMC Ecology
- BMC Evolutionary Biology
- BMJ
- BMJ Open
- Ecological Applications
- Ecological Monographs
- Ecology
- Ecosphere
- Evolution
- Evolutionary Applications
- Frontiers in Ecology and the Environment
- Functional Ecology
- Genetics
- Heredity

…

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

PLoS
PUBLIC LIBRARY of SCIENCE

JOURNAL OF Evolutionary Biology

THE ROYAL SOCIETY

Journal of Ecology
Journal of Applied Ecology
Methods in Ecology and Evolution
Functional Ecology
Journal of Animal Ecology

nature

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

http://datadryad.org/pages/jdap

The problem…

Many researchers are concerned about making their data publicly available.

This is particularly true in fields such as ecology and evolution, where datasets are often ***complex****, have a* ***long shelf life****, and can be used to* ***test multiple hypotheses****.*

Trends in Ecology & Evolution

CellPress

Opinion

Archiving Primary Data: Solutions for Long-Term Studies

James A. Mills,[1,#,*] Céline Teplitsky,[2,#,*] Beatriz Arroyo,[3] Anne Charmantier,[4] Peter. H. Becker,[5] Tim R. Birkhead,[6] Pierre Bize,[7] Daniel T. Blumstein,[8] Christophe Bonenfant,[9] Stan Boutin,[10] Andrey Bushuev,[11] Emmanuelle Cam,[12] Andrew Cockburn,[13] Steeve D. Côté,[14] John C. Coulson,[15] Francis Daunt,[16] Niels J. Dingemanse,[17,18] Blandine Doligez,[9] Hugh Drummond,[19] Richard H.M. Espie,[20] Marco Festa-Bianchet,[21] Francesca Frentiu,[22] John W. Fitzpatrick,[23] Robert W. Furness,[24] Dany Garant,[21] Gilles Gauthier,[14] Peter R. Grant,[25] Michael Griesser,[26] Lars Gustafsson,[27] Bengt Hansson,[28] Michael P. Harris,[16] Frédéric Jiguet,[29] Petter Kjellander,[30] Erkki Korpimäki,[31] Charles J. Krebs,[32] Luc Lens,[33] John D.C. Linnell,[34] Matthew Low,[35] Andrew McAdam,[36] Antoni Margalida,[37] Juha Merilä,[38] Anders P. Møller,[39] Shinichi Nakagawa,[40] Jan-Åke Nilsson,[27] Ian C.T. Nisbet,[41] Arie J. van Noordwijk,[42] Daniel Oro,[43] Tomas Pärt,[35] Fanie Pelletier,[21] Jaime Potti,[44] Benoit Pujol,[13] Denis Réale,[45] Robert F. Rockwell,[46] Yan Ropert-Coudert,[47] Alexandre Roulin,[48] James S. Sedinger,[49] Jon E. Swenson,[50] Christophe Thébaud,[12] Marcel E. Visser,[42] Sarah Wanless,[16] David F. Westneat,[51] Alastair J. Wilson,[52] and Andreas Zedrosser[53]

“A key concern is that [open data] will be a disincentive both for the initiation of long-term studies, and for maintenance of ongoing studies.”

63% of PIs were against open data as currently required

41% of respondents said that they have avoided publishing in journals that require open data

53% intend to avoid publishing in journals requiring open data in the future

Data sharing trends in ecology & evolution?

Strong open data policies have had a positive effect on data deposition rates.

(Vines et al 2013 FASEB Journal, Magee et al 2014 PLOS One)

DRYAD

figshare

zenodo

Does more mean better?...

Most journals and databases don’t verify the quality of archived data beyond basic checks like ensuring that a data availability statement and a valid DOI number are provided in the paper.

(Noor et al 2006 *PLOS Biol*, Costello et al 2013 *TREE*)

PLOS | BIOLOGY 2015

PERSPECTIVE

Public Data Archiving in Ecology and Evolution: How Well Are We Doing?

Dominique G. Roche[1,2]*, Loeske E. B. Kruuk[1,3], Robert Lanfear[1,4], Sandra A. Binning[1,2]

1 Division of Evolution, Ecology and Genetics, Research School of Biology, The Australian National University, Canberra, Australian Capital Territory, Australia, **2** Éco-Éthologie, Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland, **3** Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **4** Department of Biological Sciences, Macquarie University, Sydney, Australia

How well are we doing?

Journal	Number of studies	
	2012	2013
Biology Letters	2	10
Evolution	16	13
Evolutionary Applications	3	2
Journal of Evolutionary Biology	17	10
Nature	1	0
Science	2	3
The American Naturalist	9	12

We assessed **100 non-molecular E&E studies** in journals that had adopted the Joint Data Archiving Policy (JDAP) or had a strong open data policy.

☑ **Completeness criterion**

☑ **Reusability criterion**

Data completeness score

Meets data policy requirements

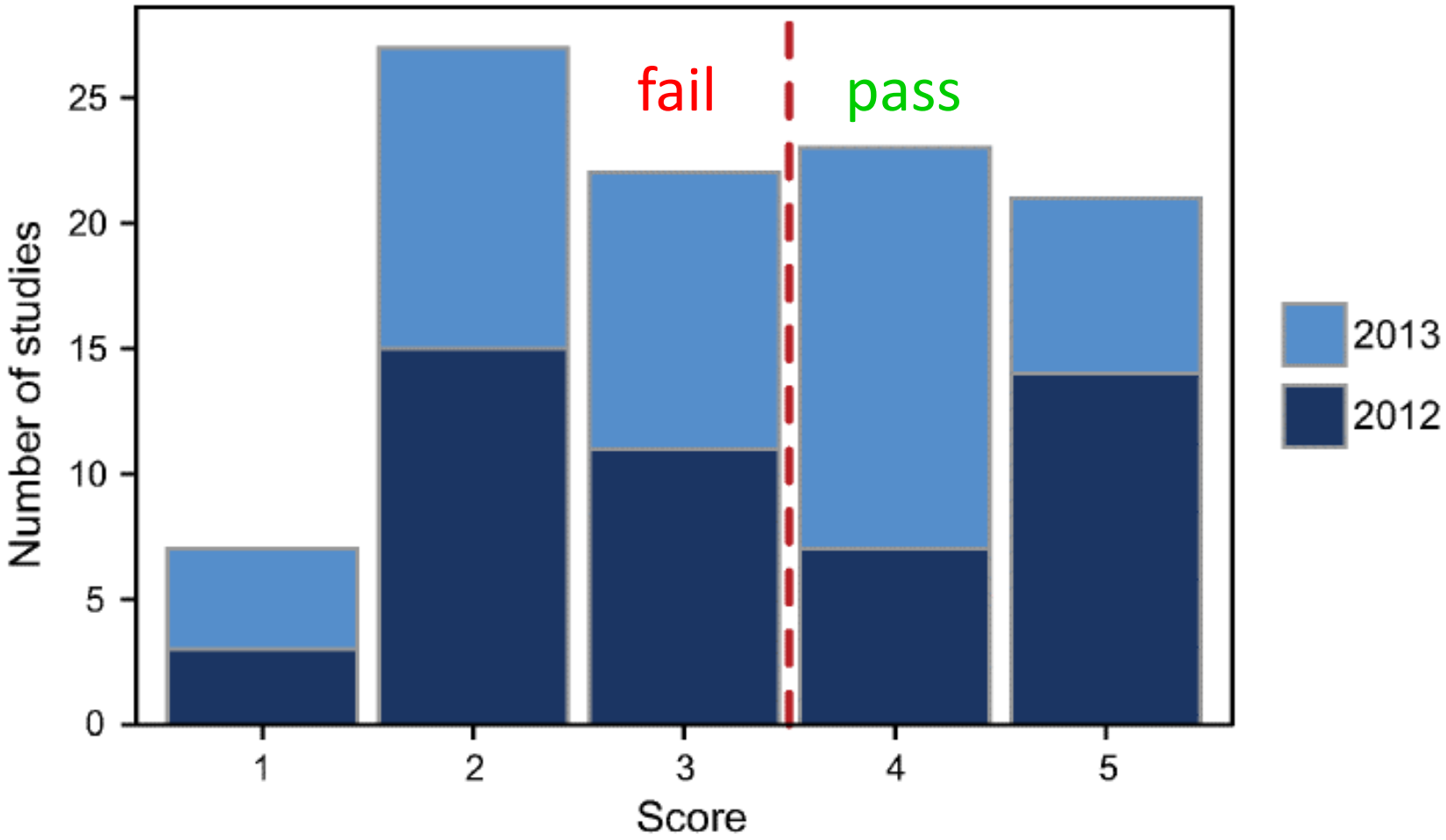
Score	Description	Criteria
5	Exemplary	All the data necessary to reproduce the analyses and results (in practice) are archived. There is informative metadata with a legend detailing column headers, abbreviations, and units.
4	Good	All the data necessary to reproduce the analyses and results (in practice) are archived. The metadata are limited or absent, but column headings, abbreviations, and units can be understood from reading the paper.
3	Small omission	Most of the data necessary to repeat the analyses are archived except for a small amount (e.g., for a supporting or exploratory analysis). The metadata are informative OR the archived data can be interpreted from reading the paper.
2	Large omission	The main analyses in the paper cannot be redone because essential data are missing AND/OR insufficient metadata or information in the paper precludes interpreting the data AND/OR the authors archived summary statistics (e.g., means), but not the raw data used in the analyses.
1	Poor	The data are not archived OR the wrong data are archived OR insufficient information is provided in the metadata or paper for the data to be intelligible.

Does not meet policy requirements

Data reusability score

Score	Description	Criteria
5	Exemplary	The data are archived in a nonproprietary, human- and machine-readable file format that facilitates data aggregation and can be processed with both free and proprietary software (e.g., csv, text; see Table 3). The metadata are highly informative (such that column headings, abbreviations, and units can be understood in isolation from the original paper). Raw data are presented (perhaps in combination with processed data such as means).[a]
4	Good	The data are archived in a format that is designed to be machine readable with proprietary software (e.g., Excel), and the metadata are highly informative (such that column headings, abbreviations, and units can be understood in isolation from the original paper). [OR] The data are archived in a nonproprietary, human- and machine-readable file format, and the metadata are sufficiently informative to be understood when combined with information from the associated paper. Raw data are presented (perhaps in combination with processed data such as means).[a]
3	Average	The data are archived in a format that is designed to be machine readable with proprietary software (e.g., Excel). The metadata are sufficiently informative to be understood when combined with information from the associated paper. Raw data are presented (perhaps in combination with processed data such as means).[a]
2	Poor	The data are archived in a human- but not machine-readable format. The metadata are highly informative OR sufficiently informative to be understood with information from the associated paper. Raw data are presented (perhaps in combination with processed data such as means).[a]
1	Very poor	The metadata are insufficient for the data to be intelligible even when combined with information from the associated paper AND/OR processed but not raw data are presented.[a]

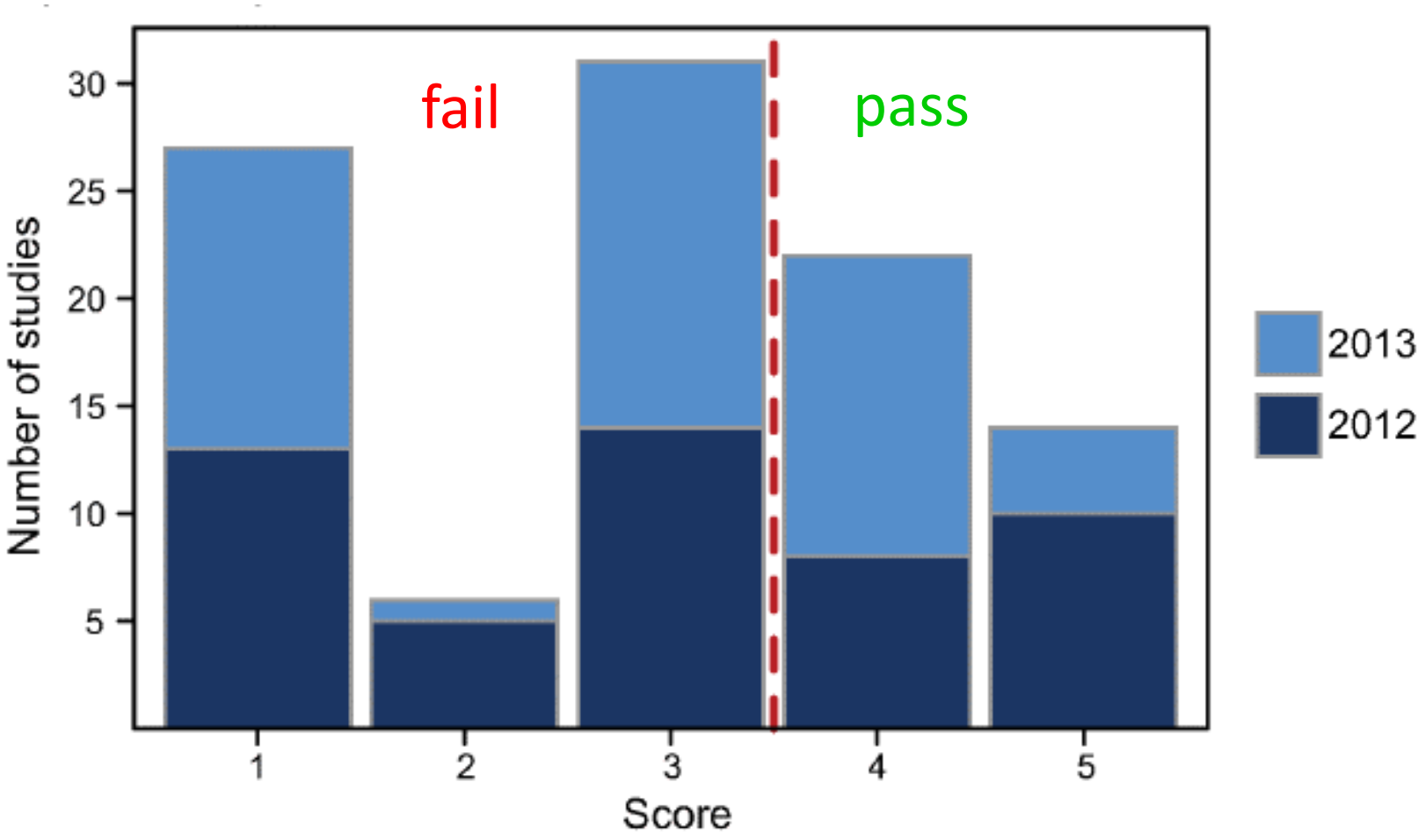
Data completeness - results



More than half (56%) of studies did not meet the minimum requirement of their journal's open data policy.

(Roche *et al* 2015 PLOS *Biol*)

Data reusability - results



<u>Even more (64%)</u> of studies were archived in a way that partially or entirely prevented reuse.

(Roche *et al* 2015 PLOS *Biol*)

How do we increase quality participation?

OPEN ACCESS Freely available online

PLOS BIOLOGY

Perspective

Troubleshooting Public Data Archiving: Suggestions to Increase Participation

Dominique G. Roche[1,2]*, Robert Lanfear[1], Sandra A. Binning[1,2], Tonya M. Haff[1], Lisa E. Schwanz[1], Kristal E. Cain[1], Hanna Kokko[1], Michael D. Jennions[1], Loeske E. B. Kruuk[1,3]

1 Division of Evolution, Ecology and Genetics, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia, **2** Australian Research Council Centre of Excellence for Coral Reef Studies, Australian National University, Canberra, Australian Capital Territory, Australia, **3** Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom








Our latest findings:

Aurelia Green

Nicholas Raihani

Jenny McClung

Redouan Bshary

UCL

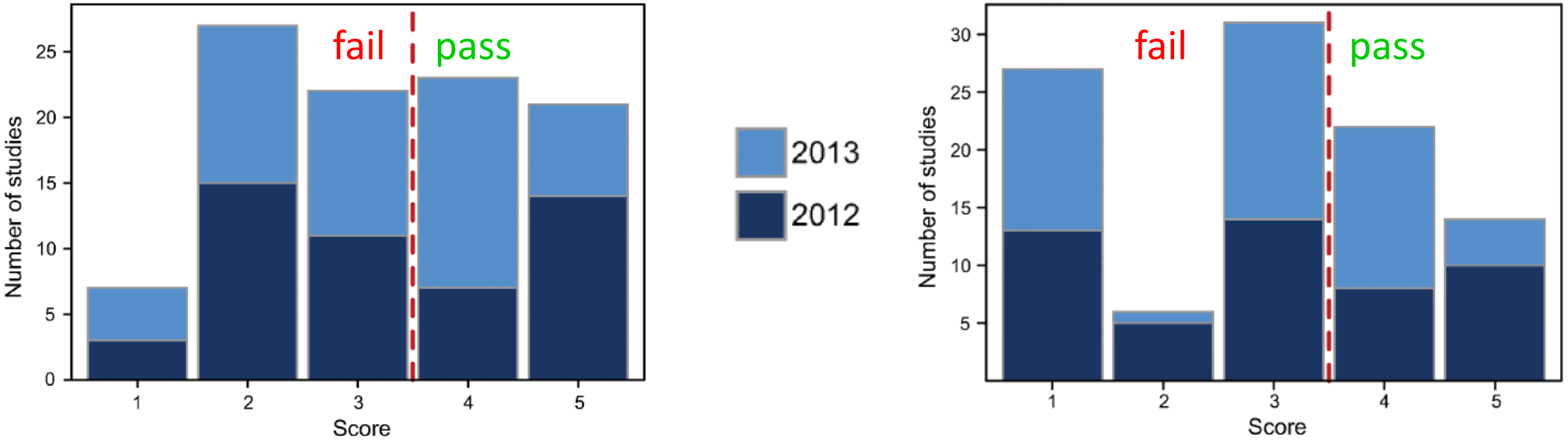
Experimental Psychology

unine

UNIVERSITÉ DE NEUCHATEL

Cognitive Science Center

Data completeness & reusability



<u>Variation</u> in the quality of open data across studies **<u>(N=100)</u>**.

What factors explain this variance?

Data sharing is perceived as ‘scholarly altruism’.

(e.g. Kim and Stanton 2015)

Humans display a ‘cooperative phenotype’ that is domain general and temporally stable

Alexander Peysakhovich[1], Martin A. Nowak[2,3,4] & David G. Rand[1,5,6]

Do cooperative scientists share more/better?

Measured cooperation of authors who shared their data as:

- response time to survey (help a graduate student)
- self-report altruism (common psychology questionnaire)
- charity donation (draw to win $500 prize for participating)

→ personalized survey requests sent out based on time-zone
→ mentioned author's research
→ 3 reminders sent at 1 week intervals

52 / 100 researchers responded to the survey

OSF

PREREGISTRATION

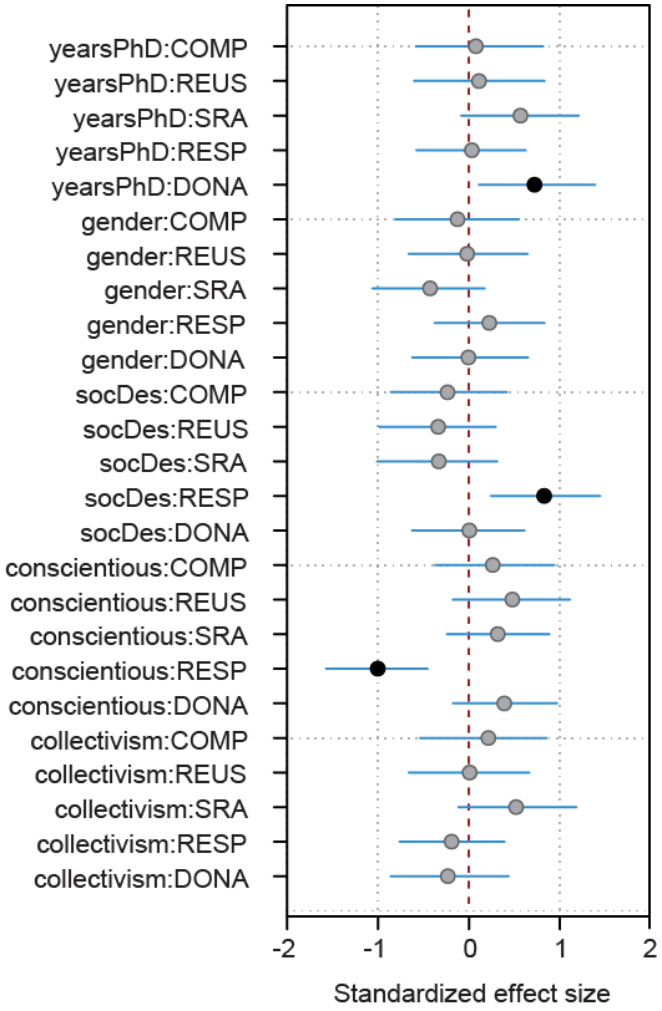
CHALLENGE

Multivariate Bayesian model to look at trait covariation

Fixed effects:

- gender
- years out of PhD
- social desirability
- conscientiousness
- collectivism

Multivariate Bayesian model to look at trait covariation

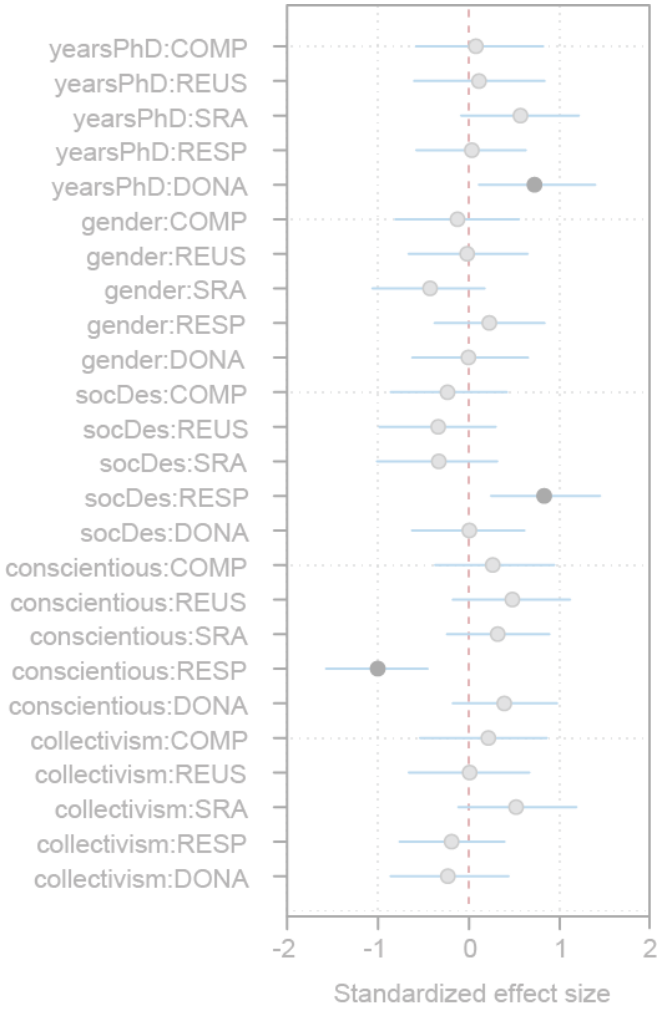


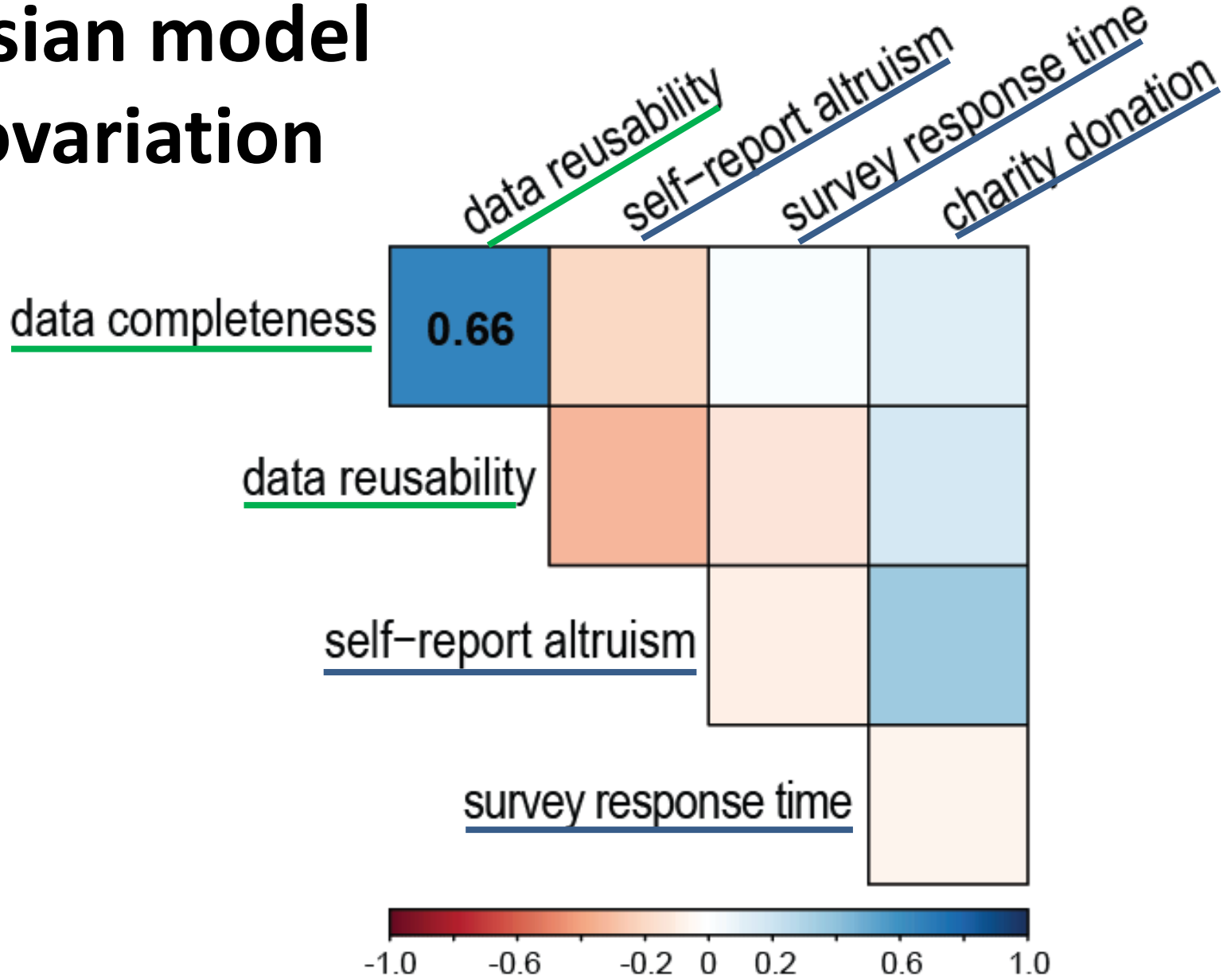
Fixed effects:

… not much happening:

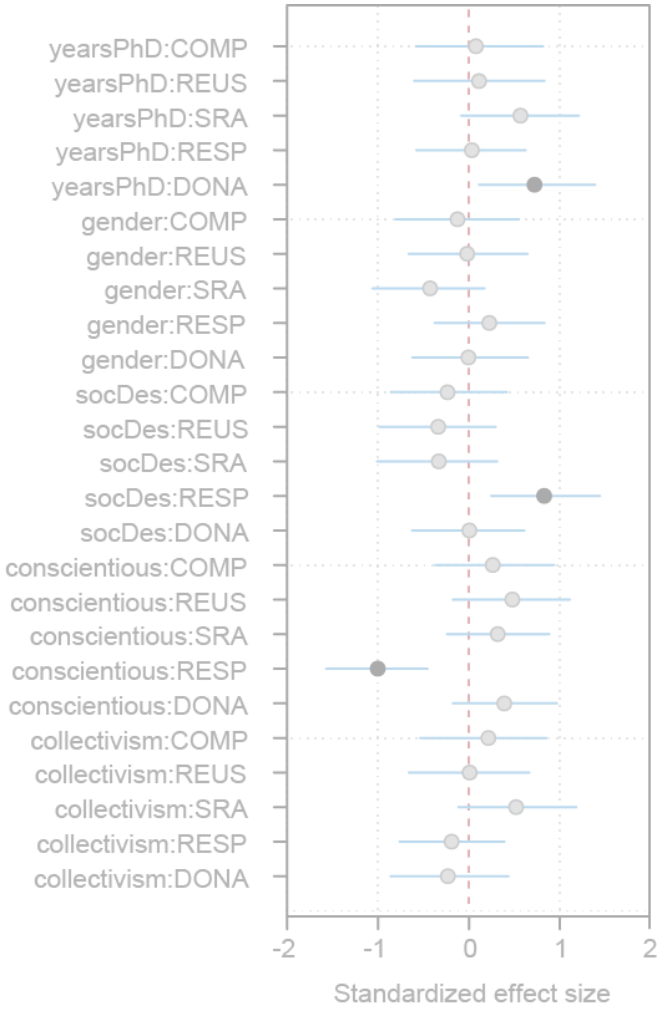
- donations ↑ with years out of PhD
- Response time ↓ with conscientiousness
- Response time ↑ social desirability

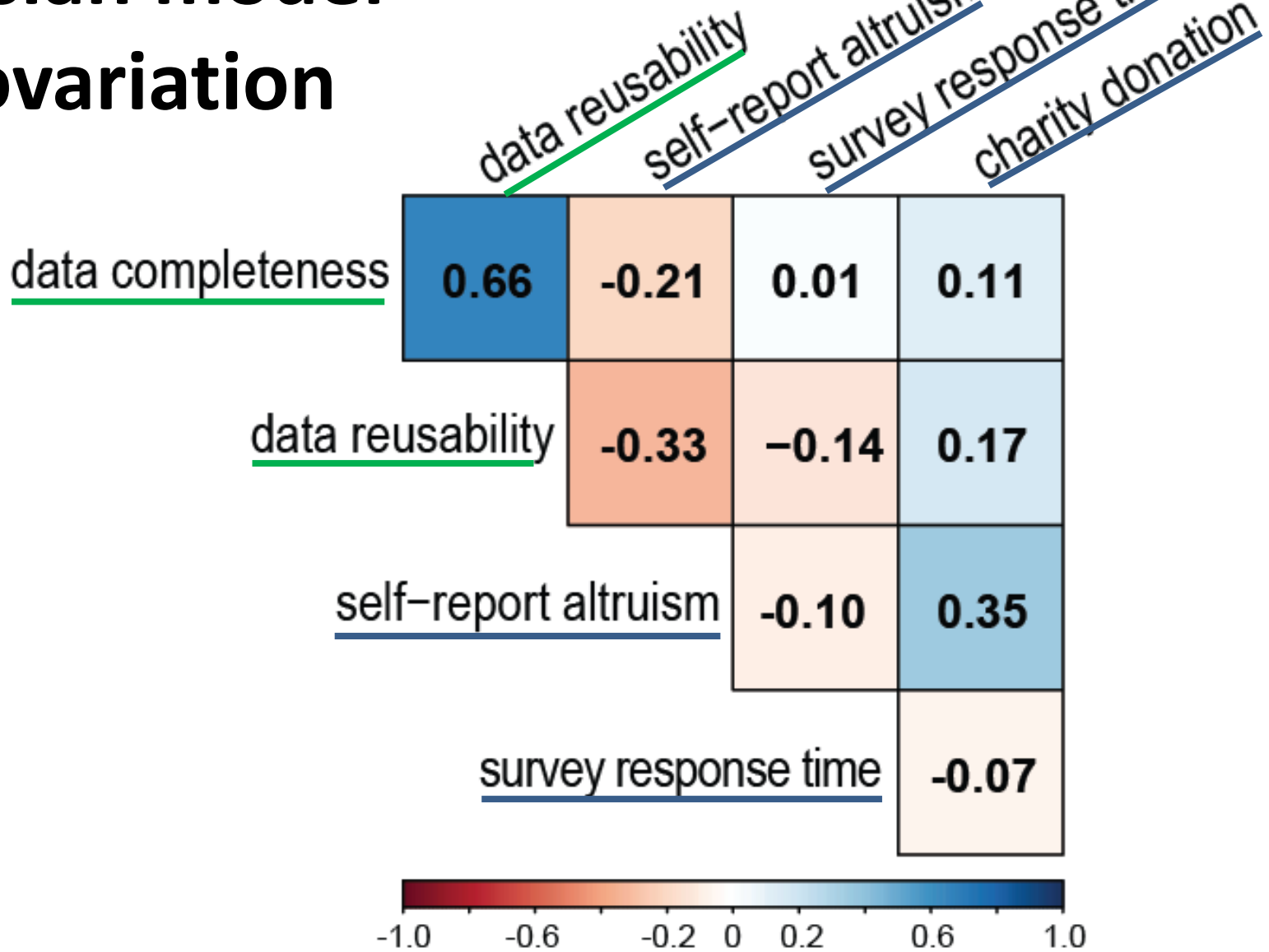
Multivariate Bayesian model to look at trait covariation





Multivariate Bayesian model to look at trait covariation





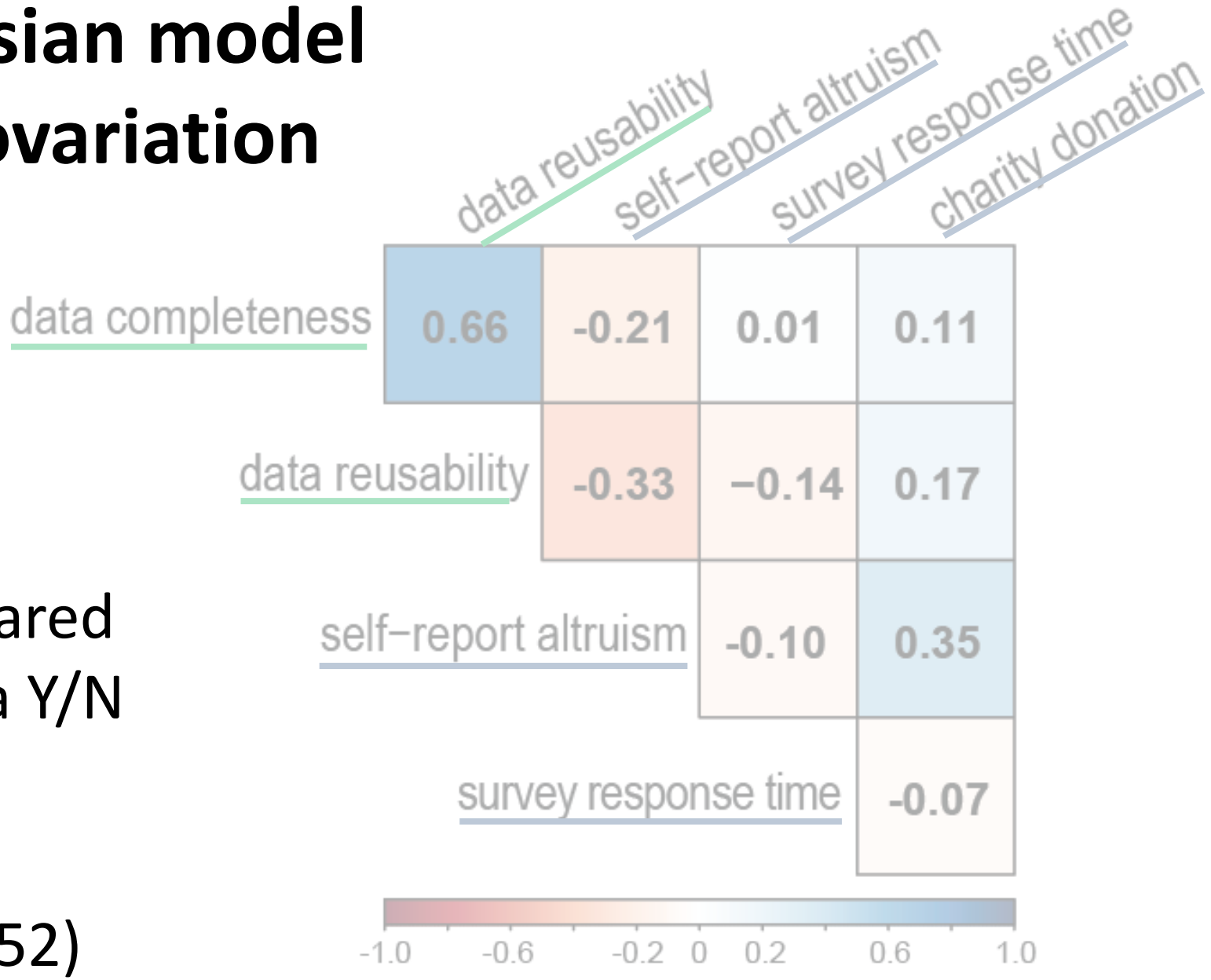
Multivariate Bayesian model to look at trait covariation

<u>Strengths:</u>

- real-world instances of cooperation
- focus on quality of shared data versus open data Y/N

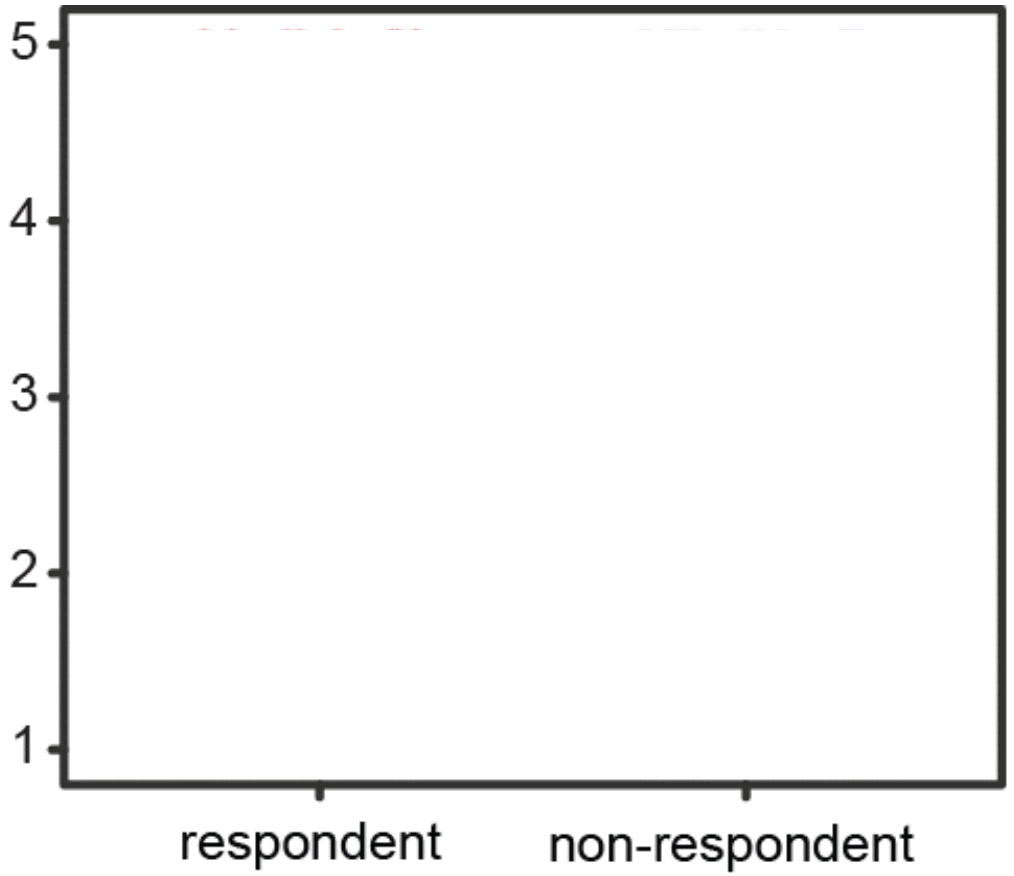
<u>Limitations:</u>

- small sample size (N=52)

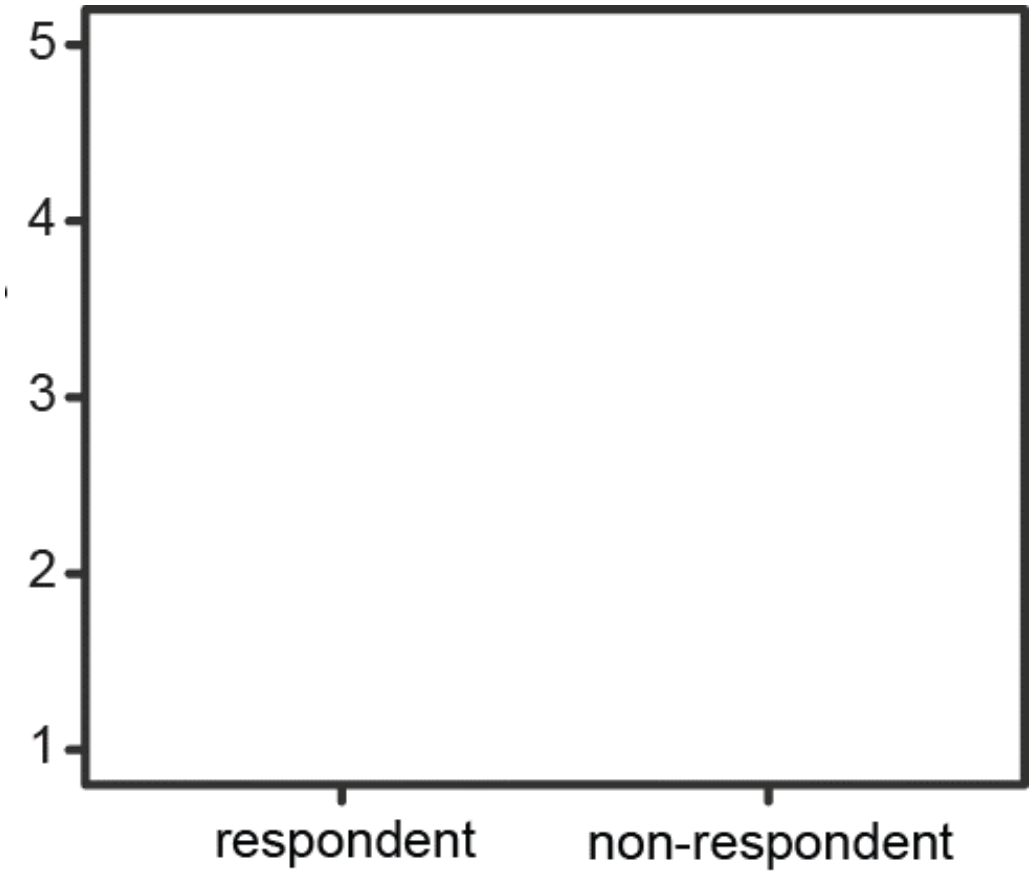


A simpler analysis… with N=100

Data completeness



Data reusability



Nice people can't share…!

We need better training in data management!

Image / illustration credits: A. Saego, Google Images

Thanks to Sandra Binning, Redouan Bshary, Loeske Kruuk, Michael Jennions, Rob Lanfear, Hanna Kokko and the eco-ethology lab at UniNE for insightful discussions.

@dom_roche

COS CENTER FOR OPEN SCIENCE
NSERC CRSNG
Carleton UNIVERSITY Canada's Capital University
unine UNIVERSITÉ DE NEUCHÂTEL
Fonds de recherche Nature et technologies Québec
figshare credit for all your research