Quels flux de recherche pour une science ouverte? Les petits ruisseaux font les grandes données

sabrina.granger@u-bordeaux.fr
Unité régionale de formation à l'information scientifique et technique (Urfist) de Bordeaux
Collège « Compétences et formation »
Comité pour la science ouverte

03/04/2019 - BootCamp #6 association VP-NUM

« While the open access movement generally focuses on the final output, we think there are also substantial benefits to an open research process. »

(Beaulieu-Jones and Greene 2017)

Ouvrir ne suffit pas

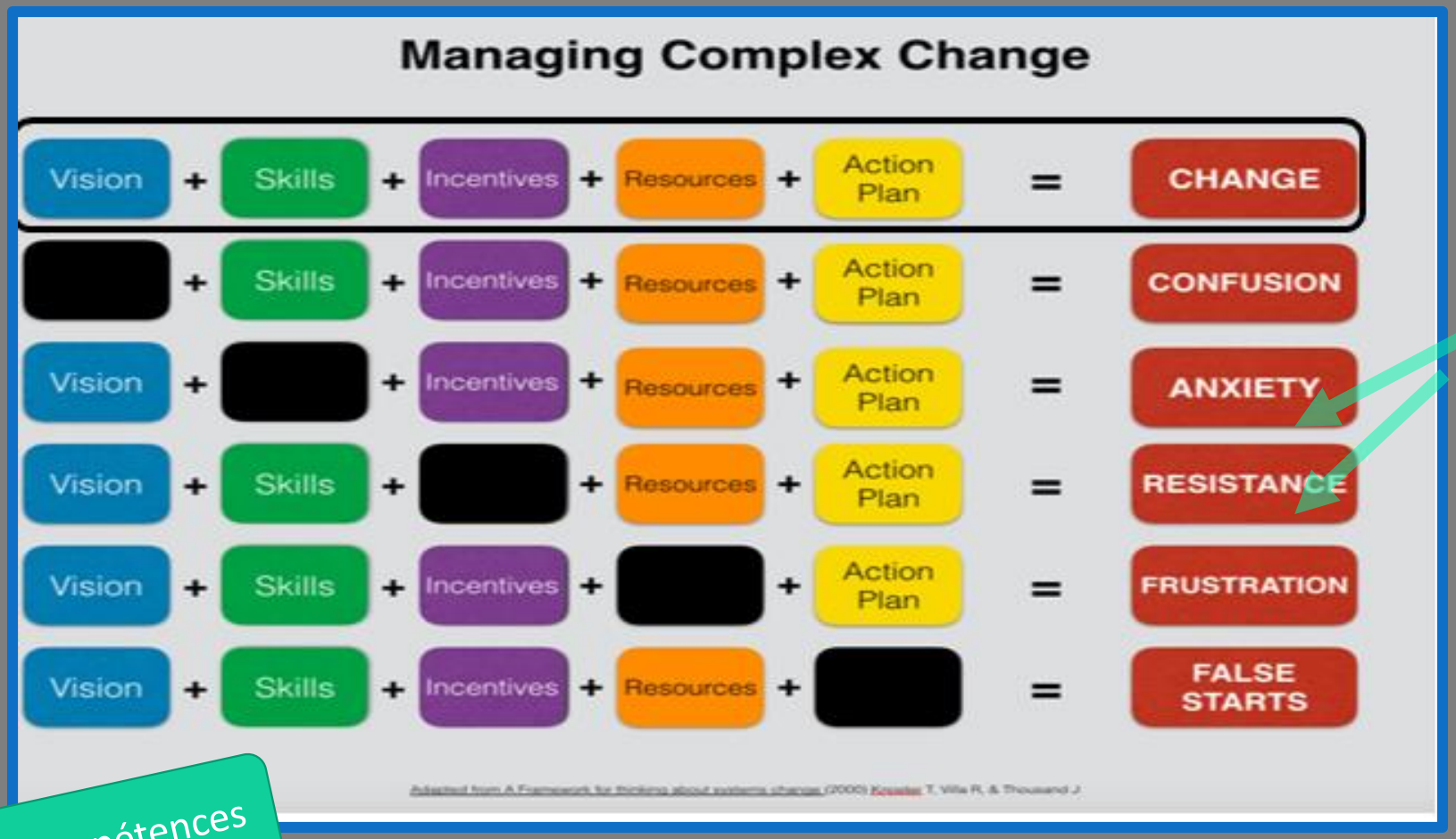
Science ouverte en action
=
Données brutes
+ Métadonnées
+ Code pour analyser les données
(Rouder 2016)

« Certainly, if the code is not available, accusations that the science cannot be trusted are easy to make.
But in reality, releasing the code makes little difference, as all but the simplest codes are impenetrable to non-experts. »
(Easterbrook 2014)

« Of the 56 articles that were then deemed potentially reproducible, we randomly chose 22 to attempt replication, and all but
1 of the 22 provided enough information that we were able to reproduce their computational findings
(given sufficient resources and a willingness write some code). » (Stodden, Seiler, and Ma 2018)

Des outils inadaptés? Ex. notebooks (Rule, Tabard, and Hollan 2018)

Vision / Skills/ Incentives / Resources / Action Plan



Compétences

Leviers

Culture(s) professionnelle(s)

Adapted from a Framework for thinking about systems changes (Knoster, Villa, Thousand, 2000) (Mendez 2019)

Compétences numériques à l'ère de l'hégémonie des données

L'acquisition des compétences : paradoxe #1

Challenges de la « data driven research »

Ecosystèmes numériques de recherche

Paradoxe #1 : ce qui est considéré comme essentiel n'est pas l'objet d'un apprentissage

Une approche empirique du numérique

(Rule, Tabard, and Hollan 2018 ; Wilson et al. 2017)

- 2 postures (contradictoires) sur le numérique #FormationBingo #AntinomieFlavour
- Place et impacts de l'informel

→ Des données conçues sans perspective sur le long terme (Lowndes et al. 2017) :

- Ni pour soi
- Ni pour les autres

Challenges de la « data driven research » (Kingsley, 2016)

Evolution des volumes de données à traiter

Evolution de la nature des données :

- Traitement et qualité des données (Leonelli 2018)

Focus sur le recours aux méthodes statistiques (Leek 2014)

→ Analyse de données comme processus exploratoire et itératif : compétences et "métacompétences" ; la donnée et son contexte

« A related set of issues is tied to **a general lack of methodological standardization**:
while many laboratories use similar methods,
the precise setups vary,
and there are few independent estimates of reliability or validity across laboratories […].»
(Frank et al., 2017, p. 423)

Ecosystèmes numériques de recherche

Une objectivation problématique des traitements numériques (Plesser 2018 ; Randall and Welser 2018)

- Utilisation au long cours (Wilson et al. 2014)
- La « tyrannie du choix » (Renata Salecl) ; ex. R et Python
- Des gouvernances hétérogènes --> question technique ET culturelle
- Le problème des logiciels propriétaires (Vihinen 2015)

« Scientific communication relies on evidence that cannot be entirely included in publications,
but the rise of computational science has added a new layer of inaccessibility.
Although it is now accepted that data should be made available on request, the current regulations regarding the availability of software are inconsistent. » (Ince et al., 2012)

Quelles incitations en faveur d'une recherche plus transparente?

Formes éditoriales émergentes et système d'évaluation : paradoxe #2

Défis techniques et culturels

Reproductibilité, répétabilité. Au-delà de la terminologie

Paradoxe #2 : Ce qui est considéré comme la clé de voûte d'une recherche transparente n'est pas une activité récompensée

« Our culture prizes innovation above all else »
(Kovačević 2007)

« We also found that gross inconsistencies are more prevalent in p-values reported as significant than in p-values reported as nonsignificant. This could suggest **a systematic bias favoring significant results**, potentially leading to **an excess of false positives in the literature**. »
(Nuijten et al., 2016)

Des défis techniques ET culturels

- Consignes aux auteurs floues (Vasilevsky et al. 2017)
- Un outillage éditorial inadapté (Allison et al. 2016) :
  - Techniquement
  - Méthodologiquement: rétractation, auto-rétractation
- « File drawer effect » (Randall and Welser 2018)
- Effet Protée (Ioannidis and Trikalinos 2005)

Reproductibilité, répétabilité : « lost in translation »

ture: authors either, *A*—make no distinction between the words *reproduce* and *replicate*, or *B*—use them distinctly. If *B*, then they are commonly divided in two camps. In a spectrum of concerns that starts at a minimum standard of "same data+same methods=same results," to "new data and/or new methods in an independent study=same findings," group 1 calls the minimum standard *reproduce*, while group 2 calls it *replicate*. *A* includes King (1995). *B*1 corresponds to the Claerbout/Donoho/Peng convention, while *B*2 agrees with Drummond (2009) and the ACM terminology. Table 2 classifies into these groups all the references cited above, and more.

Table 1: Catalogue of terminologies in the literature, with Google Scholar citations (checked Jan. 20, 2018).

A	*B*1	*B*2
King (1995), 527	Peng et al. (2006), 177	Drummond (2009), 135
JCGM (2008), 32	Gentleman and Temple Lang (2007), 216	Casadevall and Fang (2010), 58
	Laine et al. (2007), 134	Stodden (2011), 30
Dewald et al. (1986), 506	Vandewalle et al. (2009), 266	Davison (2012), 80
Pesaran (2003), 12	LeVeque (2009), 32	Loscalzo (2012), 31
McCullough et al. (2008), 93	Hyndman (2010), 20	LeVeque et al. (2012), 74
Garijo et al. (2013), 52	Jasny et al. (2011), 180	Crook et al. (2013), 16
Open Science Collaboration (2012), 300	Peng (2011), 552	Cooper et al. (2015), 26
Open Science Collaboration (2015), 1573		
Stodden (2015), 19	Koenker and Zeileis (2009), 58	Cartwright (1991), 81
Duvendack et al. (2017), 13	Delescluse et al. (2012), 22	Pellizzari et al. (2017)
Lejaeghere et al. (2016),199	Sandve et al. (2013), 227	FASEB (2016)
Coudert (2017), 3	Stodden et al. (2014), 119	
	Topalidou et al. (2015), 14	
	Iqbal et al. (2016), 67	
	Kafkafi et al. (2016), 2	
	Stevens (2017), 1	
	Kitzes et al. (2017), 10	
	Benureau and Rougier (2017), 1	
	Bollen et al. (2015), 12	
	Broman et al. (2017), 4	

Impacts des enjeux terminologiques

(Barba 2018)

Vers un nouvel ethos? Quelques éléments culturels

Transparence et autorité : paradoxe #3

Science ouverte et liberté scientifique

Paradoxe #3 : la transparence est un idéal scientifique, mais l'argument d'autorité demeure prédominant

« The credibility of **scientific claims** is rooted in the evidence supporting them, which includes the **methodology** applied, the **data** acquired, and the **process of methodology implementation**, **data analysis** and **outcome interpretation**. […]

However, without transparency, claims only achieve **credibility** based on trust in the confidence or authority of the originator.

Transparency is superior to trust.

[…] Transparency is a scientific ideal, and adding 'open' should therefore be redundant.

In reality, science often lacks openness […].» (Munafò et al., 2017)

La science ouverte, entrave à la liberté scientifique? (1/2)

Un long chemin à parcourir …

- Accusations de « terrorisme scientifique » :
  - controverse Statcheck
  - perception de la pré-registration : Scott, Sophie. 2013. 'Pre-registration would put science in chains'. Times Higher Education (THE). 25 July 2013. https://www.timeshighereducation.com/fr/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article.
- Biais cognitifs : « groupthink » (Randall and Welser 2018) ; « statistical rituals » (Gigerenzer 2018)

La science ouverte, entrave à la liberté scientifique? (2/2)

… mais une évolution déjà en cours

- Place de l'erreur et changement de perspective (Bishop 2018)
- « Adversarial collaboration » (Zwaan et al. 2017)

(Rohrer et al. 2018)

Abstract

Scientific self-correction is often construed as an outcome of the activities of the community as a whole. In contrast, cases in which researchers publicly point out errors in their own studies are rare and deemed unusual. Here, we argue that such individual self-corrections would be beneficial for the scientific community. In an online project, we invited researchers to submit statements describing how they have lost confidence in a finding they had previously published. We discuss questions that the Loss-of-Confidence project raises and argue that cultural norms should be transformed so that they live up to the pursuit of self-correction. Researchers who are interested in joining this project are invited to submit their own statement via https://lossofconfidence.com/ .

Sélection d'initiatives pour une recherche plus ouverte et plus transparente

Liste non exhaustive de solutions et de projets

(En savoir + : liens vers les sites inclus dans les logos)

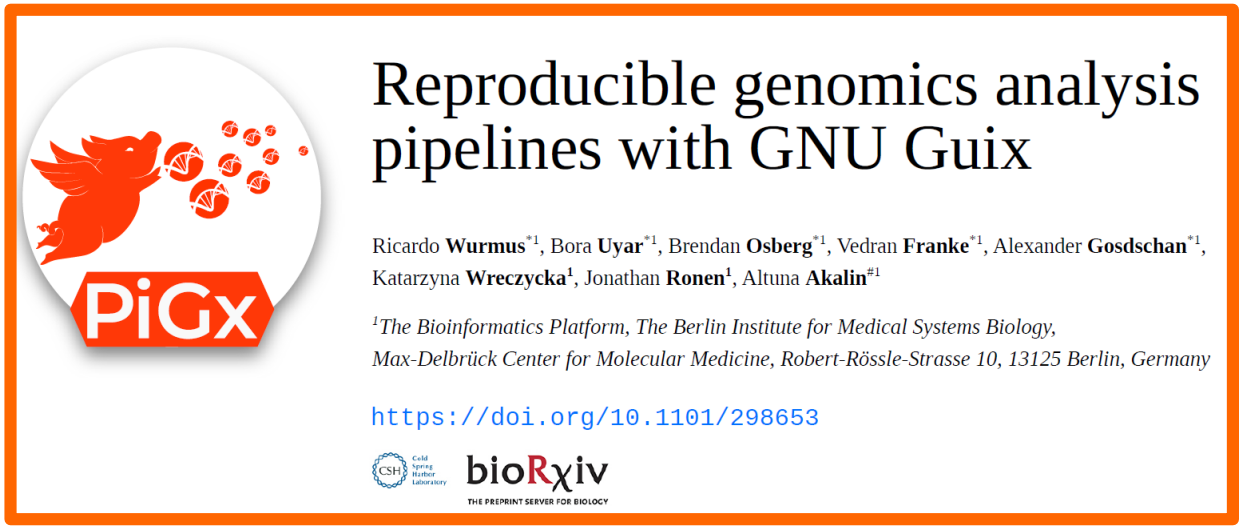
La place des méthodes et la prise de distance avec les outils :

retour aux bases des stats

Made by Sacha Epskamp and Adela Isvoranu

La portabilité des environnements logiciels :

GNU Guix, un gestionnaire de paquets (pour une stratégie complète : Guix + Software Heritage)



« Au-delà des conteneurs. Environnements logiciels reproductibles avec GNU Guix »
2018, ludovic.courtes@inria.fr (développeur Guix)

Assurer la qualité des outils pour garantir celle des données :

peer review de packages R ROpenSci

Critères de revue

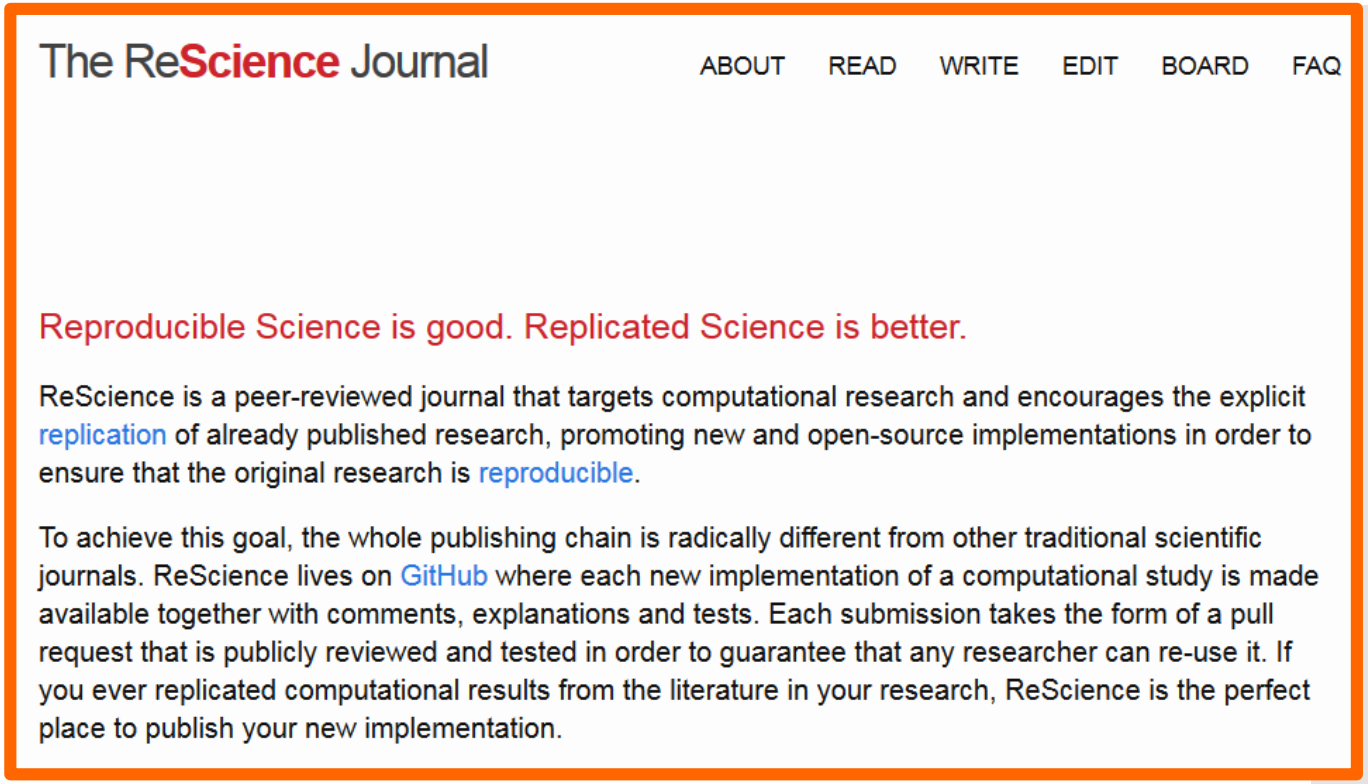
- Licence compatible avec l'*open-source initiative (OSI)*
- Documentation complète
- Haut taux de couverture de test
- Code lisible
- Utilisabilité

Tout un livre ! https://ropensci.github.io/dev_guide

maelle.salmon@yahoo.se
@ma_salmon

De nouveaux modèles éditoriaux :

structure éditoriale ad hoc pour des études de réplication



ReScience. Mieux qu'une recherche reproductible? une recherche répliquée (Reproducible Science is good. Replicated Science is better)

Nicolas Rougier, chercheur en neurosciences membre de l'équipe projet Mnemosyne

Une politique d'établissement :

OS = critère de recrutement du dpt de psychologie de Münster

The **Faculty of Humanities, Social Sciences, and Theology** invites applications for a professorship in the **Department of Psychology**. We seek to appoint an expert who will develop an internationally visible research and teaching profile in the field of sensory sciences (sensation and perception, and physiology of perception). A particular focus may lie on processes of multisensory perception and communication, taking chemo-sensory processes into account. We welcome applications from individuals with an excellent track record in research and publications on multi-sensory perception and communication. A university degree in psychology or another relevant subject from the field of life sciences is a prerequisite.

The Department of Psychology encourages transparent and replicable research and pursues these objectives with open data, open material and pre-registration. Please state clearly in your application how you have followed these objectives in the past and how you intend to do so in the future.

https://osf.io/bm5ya/

Une politique d'établissement :

pour une approche transverses des questions de méthodes

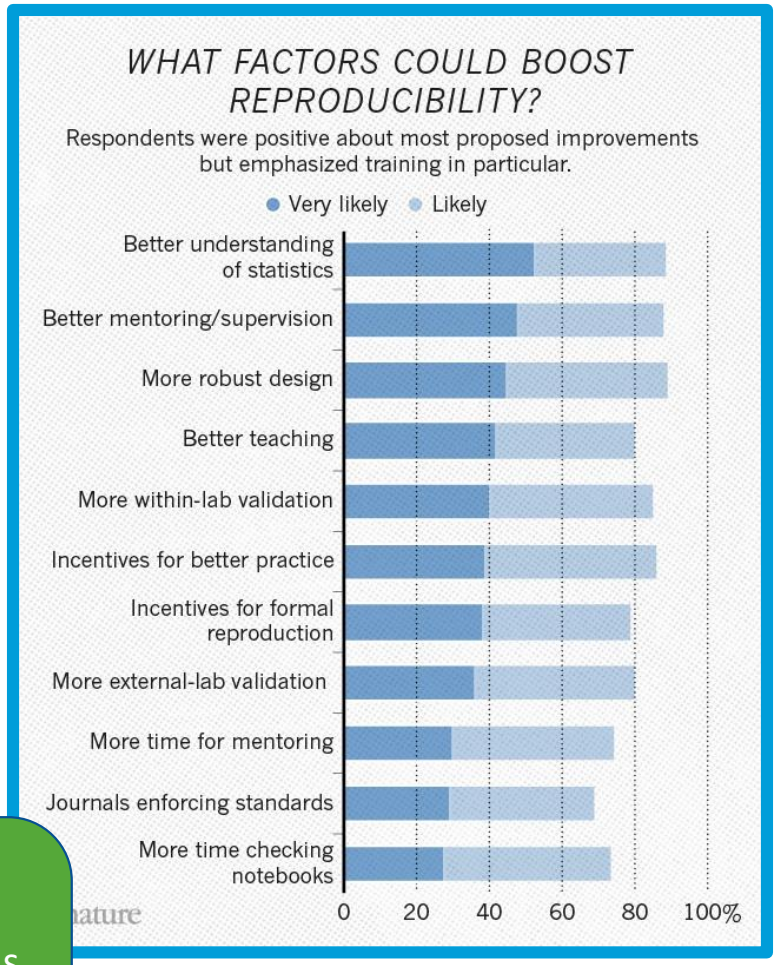
Créer des départements dédiés aux questions de méthode :

- Épistémologie
- Théorie des systèmes complexes
- Théories qui sous-tendent les statistiques
- Méthodologie de data sharing
- Méthodologie de la réplication
- Study design, experimental design …

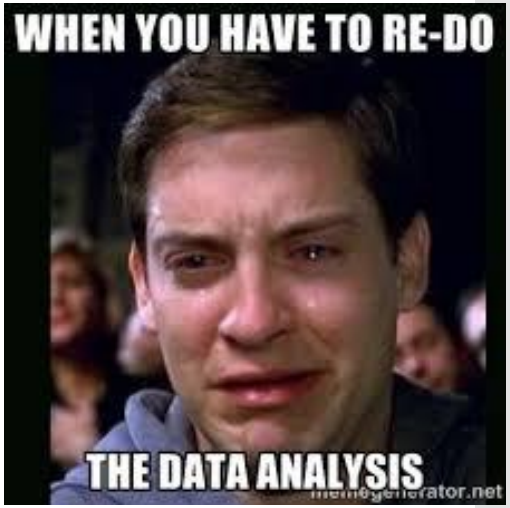
(Randall and Welser 2018)

Dpt of Methodology and Statistics, Tilburg University

1 feuille de route conçue par des chercheur.euse.s



(Baker 2016)



\+

Favoriser le travail collaboratif et les occasions d'échanges entre disciplines
Aider les chercheur.euse.s à documenter leurs workflows
…

Pour la route : un classique

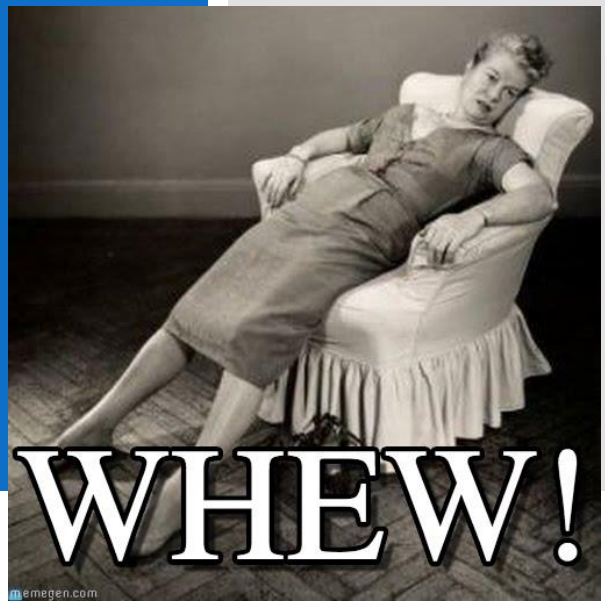
Problem 1: Data Excuse Bingo

My data contains personal/sensitive information	My data is too complicated	People may misinterpret my data	My data is not very interesting
Commercial funder doesn't want to share it	We might want to use it in another paper	People will contact me to ask about stuff	Data Protection/ National Security
It's too big	People will see that my data is bad	I want to patent my discovery	It's not a priority and I'm busy
I don't know how	I'm not sure I own the data	Someone might steal/ plagiarise it	My funder doesn't require it

Data Excuse Bingo created by @jenny_molloy

Merci pour votre écoute

sabrina.granger@u-bordeaux.fr



Sources
(powered by Zotero)

Allison, David B., Andrew W. Brown, Brandon J. George, and Kathryn A. Kaiser. 2016. 'Reproducibility: A Tragedy of Errors'. *Nature News* 530 (7588): 27. https://doi.org/10.1038/530027a.

Baker, Monya. 2016. '1,500 Scientists Lift the Lid on Reproducibility'. Nature News. 26 May 2016. http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

Barba, Lorena A. 2018. 'Terminologies for Reproducible Research'. *ArXiv:1802.03311 [Cs]*, February. http://arxiv.org/abs/1802.03311.

Beaulieu-Jones, Brett, and Casey Greene. 2017. 'Reproducibility: Automated. Repurposing Continuous Integration Tools for Scientific Analyses Takes the Headache out of Reproducible Research'. *ELife*, October. https://elifesciences.org/labs/e623676c/reproducibility-automated.

Bishop, D. V. M. 2018. 'Fallibility in Science: Responding to Errors in the Work of Oneself and Others': *Advances in Methods and Practices in Psychological Science*, July. https://doi.org/10.1177/2515245918776632.

Easterbrook, Steve M. 2014. 'Open Code for Open Science?' *Nature Geoscience* 7 (11): 779–81. https://doi.org/10.1038/ngeo2283.

Frank, Michael C., Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, et al. 2017. 'A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building'. *Infancy* 22 (4): 421–35. https://doi.org/10.1111/infa.12182.

zotero

Sources
(powered by Zotero)

Gigerenzer, Gerd. 2018. 'Statistical Rituals: The Replication Delusion and How We Got There, Statistical Rituals: The Replication Delusion and How We Got There'. *Advances in Methods and Practices in Psychological Science* 1 (2): 198–218. https://doi.org/10.1177/2515245918771329.

Ince, Darrel C., Leslie Hatton, and John Graham-Cumming. 2012. 'The Case for Open Computer Programs'. *Nature* 482 (7386): 485–88. https://doi.org/10.1038/nature10836.

Ioannidis, John P. A., and Thomas A. Trikalinos. 2005. 'Early Extreme Contradictory Estimates May Appear in Published Research: The Proteus Phenomenon in Molecular Genetics Research and Randomized Trials'. Journal of Clinical Epidemiology 58 (6): 543–49. https://doi.org/10.1016/j.jclinepi.2004.10.019.

Kingsley, Danny. 2016. 'Reward, Reproducibility and Recognition in Research – the Case for Going Open'. Munin, November 21. http://dx.doi.org/10.7557/5.4036.

Kovačević, Jelena. 2007. 'How to Encourage and Publish Reproducible Research'. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* IV: 1273–76.

Leek, Jeff. 2014. 'On the Scalability of Statistical Procedures: Why the p-Value Bashers Just Don't Get It. · Simply Statistics'. *Simply Statistics* (blog). 14 February 2014. https://simplystatistics.org/2014/02/14/on-the-scalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/.

Leonelli, Sabina. 2018. 'Re-Thinking Reproducibility as a Criterion for Research Quality'. *History of Economic Thought and Methodology*, January, 19.

Sources
(powered by Zotero)

Lowndes, Julia S. Stewart, Benjamin D. Best, Courtney Scarborough, Jamie C. Afflerbach, Melanie R. Frazier, Casey C. O'Hara, Ning Jiang, and Benjamin S. Halpern. 2017. 'Our Path to Better Science in Less Time Using Open Data Science Tools'. *Nature Ecology & Evolution* 1 (6): 0160. https://doi.org/10.1038/s41559-017-0160.

Mendez, Eva. 2019. 'Open Science?… Darling, We Need to Talk!' presented at the Open Science Conference 2019, Berlin, March 18.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. 'A Manifesto for Reproducible Science'. *Nature Human Behaviour* 1 (1): 0021. https://doi.org/10.1038/s41562-016-0021.

Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. 'The Prevalence of Statistical Reporting Errors in Psychology (1985–2013)'. *Behavior Research Methods* 48 (4): 1205–26. https://doi.org/10.3758/s13428-015-0664-2.

zotero

Sources
(powered by Zotero)

Plesser, Hans E. 2018. 'Reproducibility vs. Replicability: A Brief History of a Confused Terminology'. *Frontiers in Neuroinformatics* 11. https://doi.org/10.3389/fninf.2017.00076.

https://doi.org/10.1371/journal.pcbi.1005510.

Randall, David, and Christopher Welser. 2018. *The Irreproducibility Crisis of Modern Science. Causes, Consequences, and the Road to Reform*. New York: National Association of Scholars.

Rohrer, Julia Marie, Lisa Marie DeBruine, Tom Heyman, Benedict C Jones, Stefan C. Schmukle, Raphael Silberzahn, Eric Luis Uhlmann, Rebecca Magdalena Willén, Christopher Chabris, and Tal Yarkoni. 2018. 'Putting the Self in Self-Correction'. *PsyArXiv*, December. https://doi.org/10.31234/osf.io/exmb2.

Rouder, Jeffrey N. 2016. 'The What, Why, and How of Born-Open Data'. *Behavior Research Methods* 48 (3): 1062–69. https://doi.org/10.3758/s13428-015-0630-z.

Sources
(powered by Zotero)

Rule, Adam, Aurélien Tabard, and James D Hollan. 2018. 'Exploration and Explanation in Computational Notebooks'. In . Montréal, Canada. https://hal.archives-ouvertes.fr/hal-01676633.

Stodden, Victoria, Jennifer Seiler, and Zhaokun Ma. 2018. 'An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility'. *Proceedings of the National Academy of Sciences* 115 (11): 2584–89. https://doi.org/10.1073/pnas.1708290115.

Vasilevsky, Nicole A., Jessica Minnier, Melissa A. Haendel, and Robin E. Champieux. 2017. 'Reproducible and Reusable Research: Are Journal Data Sharing Policies Meeting the Mark?' *PeerJ* 5 (April): e3208. https://doi.org/10.7717/peerj.3208.

zotero

Sources
(powered by Zotero)

Vihinen, Mauno. 2015. 'No More Hidden Solutions in Bioinformatics'. *Nature* 521 (7552): 261. https://doi.org/10.1038/521261a.

Wilson, Greg, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, et al. 2014. 'Best Practices for Scientific Computing'. *PLOS Biology* 12 (1): e1001745. https://doi.org/10.1371/journal.pbio.1001745.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. 'Good Enough Practices in Scientific Computing'. *PLOS Computational Biology* 13 (6): e1005510.

Zwaan, Rolf A., Alexander Etz, Richard E. Lucas, and M. Brent Donnellan. 2017. 'Making Replication Mainstream'. *Behavioral and Brain Sciences*, October, 1–50. https://doi.org/10.1017/S0140525X17001972.

zotero