

Reproducibility, what?, why?, when?

Experiences within the IEEE and ideas to move forward



Gianluca Setti

Department of Engineering, University of Ferrara, Italy
2013-2014 IEEE Vice President, Publication Services and Products

`gianluca.setti@unife.it`

Panel: **Publishing Innovations for Research reproducibility**
STM Innovation Seminars, 2015



London, UK
December 2, 2015



References

- [1] J. B. Buckheit and D. L. Donoho. (1995). "WaveLab and reproducible research," Dept. of Statistics, Stanford Univ., Tech. Rep. 474. [Online]. Available: <http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf>
- [2] <http://en.wikipedia.org/wiki/Reproducibility>
- [3] http://www.nytimes.com/2006/05/12/world/asia/12korea.html?_r=0
- [4] M. Barni, F. Perez-Gonzalez, "Pushing science into signal processing [my turn]," Signal Processing Magazine, IEEE , vol.22, no.4, pp.120,119, July 2005
- [5] P. Vandewalle, J. Kovacevic and M. Vetterli, "Reproducible Research in Signal Processing - What, Why, and How," IEEE Signal Processing Magazine, 2009
- [6] Piwowar, H. a et al. Sharing detailed research data is associated with increased citation rate. PloS one. 2, (2007), 308

- 2013-2014 IEEE Vice President for Publication Services and Products (PSPB) and as 2012 Chair of the IEEE Strategic Planning Committee of the PSP Board
 - Data repositories
 - IEEE Digital Library (*Xplore*) enhancement for storing "not only papers" (videos, code, ...)

- Chair of the 2015 ad-hoc Committee of the IEEE Board of Directors on *Future of Information & Convening*
 - Committee charter includes **promoting reproducible research**

- *Reproducible research* refers to the idea that the ultimate product of academic research is the paper **along with** the full computational environment used to produce the results in the paper **such as the code, data, etc.** that can be used to reproduce the results and create new work based on the research (<http://en.wikipedia.org/wiki/Reproducibility>)

- “An article [...] in a scientific publication is not the scholarship itself, it is the actual scholarship which is done in an environment which is not reproducible and reproducible research,” Dept. of Statistics, Stanford Univ., Tech. Rep. 474.)

Not only the paper is important!!

RR=paper+data+algorithm+code
+experimental methods

Paper, Data, Algorithm, Code, Experiments
are equally important scholarly products

- Mathematical Sciences

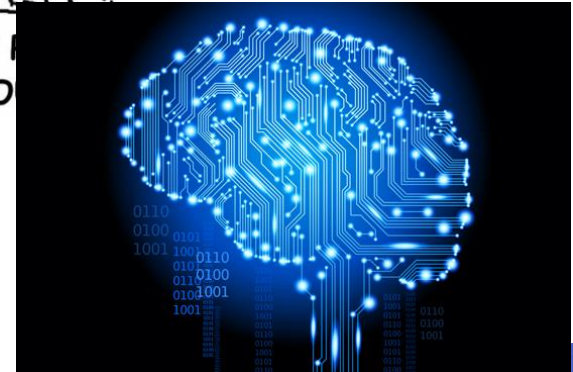
Easy



YOU WANT
I'LL GIVE YOU

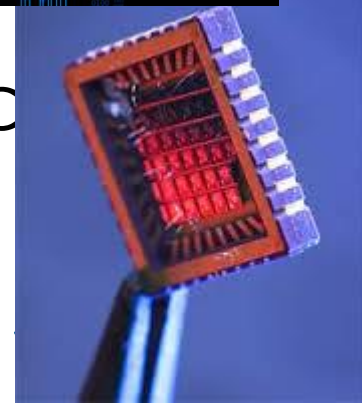
- Computational Sciences

Intermediate



- Empirical Sciences (e.g. Life Sciences, Design, System implementation)

Difficult



- IEEE is **first a professional organization** and its publishing enterprise serves as a service to the community

- Advantages of RR (for IEEE)

- Increase collaboration capabilities (obvious): science is all "about collaboration" and "about standing on the shoulders of giants"



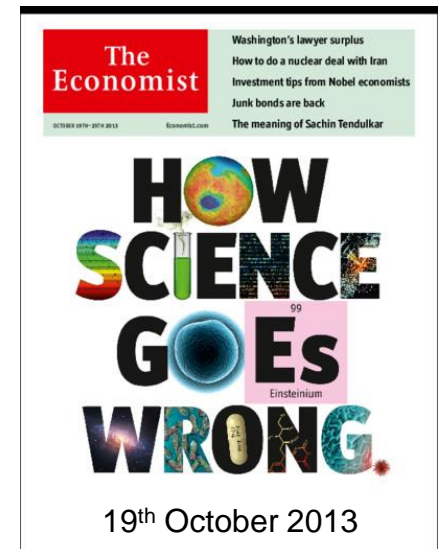
Stand on the shoulders of giants

- Information will be more visible and directly usable:
 - ✓ **Researchers will advance technology more easily**
 - ✓ **Practitioners will develop new products faster**
 - ✓ **RR can reduce the quantity of "noise" (very important in the era of information curation)**



- Make the review process more reliable (and faster?)
 - ✓ **Reviewers are a precious and non scalable resource**
- Make it more difficult (impossible?) to plagiarize a paper:
 - ✓ Unfortunately the number of plagiarism cases is rapidly increasing despite IEEE education efforts so far
 - ✓ **RR will make plagiarized research more easy to discover**
 - ✓ **Education in RR will help to establish a more uniform ethical attitude in the global community**
- Make it easier to discover false results and avoid retractions

Stem-cell scandal: Dr. Hwang Woo-suk became infamous for fabricating a series of experiments which appeared in Science in 2004 and 2005 have succeeded in creating human embryonic stem cells by cloning [3]



- This case is worse since ... the study is actually "correct", simply it was designed to produce meaningless results
- The data set is too small to be statistically significant
- It is a (ad-hoc constructed) example of **Questionable Research Practice** (as outliers eliminations, "p-hacking"), which is a waste of resources and **much more common**

tion, in par
have rarely



- Main player so far: IEEE Signal Processing Society (SPS. Area: Computational Science)
 - Barni and Perez-Gonzalez started the discussion in SPS [4] (2005)
 - Vetterli and his group championed the topic in SPS creating a web page and a repository for RR at <http://rr.epfl.ch/> [5]
 - Other prominent researchers in SPS (e.g. Kovacevic [6]) have been active in promoting RR
- Starting data point: How many papers published in the *IEEE Transactions on Signal Processing* in 2004 were reproducible?

[TABLE 1] RESULTS OF REPRODUCIBILITY STUDY ON IEEE TRANSACTIONS ON IMAGE PROCESSING PAPERS PUBLISHED IN 2004. AVERAGE SCORES OVER THE 134 PAPERS ARE PRESENTED.

ALGORITHM		CODE				DATA				
PARAMETER DETAILS	BLOCK DIAGRAM	PSEUDO-CODE	PROOFS	COMPARISON	IMPLEM. DETAILS	CODE AVAIL.	EXPLANATION OF DATA	SIZE DATA SET	DATA AVAIL.	
0.84	0.71	0.37	0.33	0.27	0.64	0.12	0.09	0.83	0.47	0.33

from [5]

Not particularly encouraging ☹️

Has RR percolated in SPS/IEEE since then?

Have we been more successful than the medical science?

NO

- All SPS publications encourage RR
 - (Instruction for Authors) "To make your work reproducible by others, the TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper"
- Most other IEEE journals do not actively promote RR in this manner
- Our digital library (IEEE Xplore) allows the addition of supplemental material to a paper, where data, algorithms and code, ... can be stored
 - No standard format followed, reproducibility still difficult

NOT YET

Has RR percolated in SPS/IEEE since then?

- Create a searchable and addressable repository for data.
 - Unique identifier is needed (DOI-like) to reference data set allowing credit to authors
 - Use of standardized set of data for specific problems should be encouraged
- Create a cloud/cloud-based repository for code/algorithms/circuits which make it easier the reuse
 - Installing and running the code should not require the authors intervention!!!
 - **Public/Private partnership may be required** (Software companies, Funding Agencies, Professional Organizations, Publishers, ...)

- Authors can upload algorithms which can be linked to the corresponding papers. **Develop metrics to track algorithm usage: algorithms and corresponding papers should be cross-linked.**

▪ The culture needs to be changed

- Our community has not made sufficient effort to reward the positive efforts of authors contributing to RR
 - ✓ **A well-prepared RR contribution requires time and effort, which may be in opposition to the *publish or perish* pressure**
 - ✓ As a publisher, we should facilitate the highlighting of papers with RR content
 - ✓ We can create awards for RR papers only
 - ✓ We can expedite the review of RR papers

- There is evidence that RR papers are more highly downloaded and cited than non-RR papers
 - With the rise of citation databases and bibliometrics, today's emphasis is not only on solely the act of publishing, but also on achieving high visibility with publications (e.g., in a study in 2007 on microarray DNA research publications papers with shared data were cited about 70% more frequently!!)

- **Can funding agencies help move RR forward?**
- **Will funding impose RR?**
 - Should/will RR be required for publicly funded research?
 - Should/will such a requirement be global (which may be difficult, given what is happening for Open Access)?
 - Should/will funding agencies participate in setting up the infrastructure required for RR?

- There is a push by an increasing part of the scientific community to move in the directions to **promote reproducible research**

NATIONAL RESEARCH COUNCIL

OF THE NATIONAL ACADEMIES

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: A Workshop

February 26-27, 2015

- Representatives of IEEE, SIAM, ACM, NIH, NSF, and several expert in the

**RR is a Collective action problem:
coordination of researcher incentives, universities,
funding agencies, journals,
scientific societies, legal and policy environment,
internal and ethical pressures, libraries, the public.**

- (At the National Academy meeting) NSFs and NIH representative stated the intent of the research agencies to **push for RR**

**Reproducible Research could
be the
"Next Government Mandate"
(Recall Open Access?)**

- Reproducible Research is known and appreciated by a specific IEEE community but has not had a large penetration
- There is value in promoting RR for the IEEE
 - in its role as a publisher
 - to allow faster research advancement in its technical communities
- IEEE is actively considering supporting RR materials (code/algorithms/data) in its digital library
 - It may require a public/private partnership
- A "procedure" is needed for RR. Should we all engage in "standards" for RR?

Questions?