HOW TO ENCOURAGE AND PUBLISH REPRODUCIBLE RESEARCH

Jelena Kovačević

Depts. of Biomedical Engineering & Electrical and Computer Engineering
Carnegie Mellon University
Email: jelenak@cmu.edu

ABSTRACT

I discuss the "what", "why" and "how" or reproducible research, a concept that emerged recently in computational sciences. It refers to the idea that the ultimate product is not a published paper only, but the data, software and everything else needed to produce that paper. In signal processing, the discussion just started, and this paper attempts to add to the current efforts of bringing the issue to the forefront and looking for solutions to make it happen.

Index Terms— Reproducible research, literate programming.

1. WHAT IS REPRODUCIBLE RESEARCH AND WHY DO WE NEED IT?

"1993: Cracking math's oldest brain-teaser" and "Cloned human embryos are stem cell breakthrough" were the headlines that excited the world. While the first one turned out to be true and one of the last century's greatest achievements, the second developed into a historic scandal when it was revealed that the data was doctored. Both of these instances bring about the issue of reproducible research: what is it, why do we do it, how do we get to reproducible research and finally, how we publish true reproducible research.

Reproducible research (RR) refers to the idea that in "computational" sciences, the ultimate product is not a published paper but rather the entire environment used to produce the results in the paper (data, software, etc.). While it might sound natural and obvious, how many of us in signal processing really do it that way?

1.1. Theory Versus Experimentation

Throughout history, scientific achievements have been roughly divided into two categories: theoretical and experimental. In either of these, the "reproducibility" was established in a specific way:

In theoretical disciplines, such as mathematics for example, abstract results—theorems—are built starting from "given truths"—axioms, on which a logical pyramid is built—a proof. Each step of the pyramid relies on the axioms as well as the other known theorems. The validity of the proof is ascertained through peer review—examination of the proof by other mathematicians. Once through that stage, the theorem is added to the set of tools used to build other theorems. The issue of reproducibility is settled at that point; the proof allows anyone to reproduce the steps leading to the theorem. While it is certainly possible that another mathematician might come up with a shorter/more elegant/easier/... proof, the theorem is already known to be true and the steps to reaching its conclusion(s) are published and reproducible.

The author's work is supported by NSF through awards CCF-0515152 and EF-0331657 as well as by the PA State Tobacco Settlement, Kamlet-Smith Bioinformatics Grant.

In experimental disciplines, such as biology for example, the reproducibility has another form. The biologist forms a hypothesis (equivalent to the mathematician's theorem, which is a hypothesis until proven), and then proceeds to prove or disprove the hypothesis by performing experiments. Thus, what mathematicians would call a proof would, in biology, be the methodology, the set of experiments as well as the resulting data and its interpretation, that would prove the hypothesis. While, when written, such works go through the same process of peer review, the result does not become a "theorem" until at least another independent group is able to perform the exact same experiments and confirm the results. Of course, to truthfully replicate the experiments, the paper has to provide enough specific detail about the experiments to allow another group to mimic it—the reproducibility criterion.

While the above criteria seem simple in theory, in practice things do not always work smoothly. For example, it is quite possible for a mathematician to make a mistake in the proof of a theorem and for this mistake to go unnoticed. In the late 1990s, it took mathematicians two years to check the Andrew Wiles' proof of Fermat's Last Theorem. The original proof had a flaw which first went unnoticed and which Wiles eventually fixed. Some not so happy examples from recent times include the stem-cell scandal; The reputed researcher Hwang Woo Suk claimed to have created tailor-made stem cell colonies which in fact did not exist. Although the results were published in one of the most famous and respectable journals—Science (which means the article went through a rigorous review process), the results were completely fraudulent.

1.2. A Hybrid Is Born: Computational Sciences

To make matters more interesting, in the second half of the last century, a "hybrid" of theoretical and experimental research developed: research in computational sciences. These sciences encompass many fields, including computer science, statistics, many areas of engineering, as well as our own—signal processing. Take signal processing as an example. It is debatable whether we have adopted the good practices from either our theoretical or our experimental ancestors. In terms of theoretical rigor, we often find our publications vague, hand-waving and with simplest experiments as proof. However, to our defense, there is a number of researchers who clearly state the assumptions and develop logical proofs. On the experimental side, the situation is bleaker; In a host of our papers, no scientific methodology is followed, no comparisons to competing techniques are given and/or sample sizes are dismally small and no confidence intervals are given.. Several of these issues were brought up by LeVeque [1] as well as Fomel and Hennenfent [2].

1.3. Birth of Reproducible Research

The awareness brought about by these issues has prompted the surge in the last decade of RR initiatives. While *reproducible research* as a term seems to have sprung up only in the 1990s, one can trace its predecessor ideas to none other than the man we all admire, Donald Knuth, under the name *literate programming*. Knuth pioneered the concept in the 1980s [3], with the premise that the "programs are useless without descriptions", "descriptions should be literate, not comments in code or typical reference manuals", "the code in the descriptions should work", and "it is necessary to extract the real working code from the literary description", as eloquently described on Greyer's site [4].

This was followed by the RR movement, with Claerbout as one of the pioneers [5]. In Claerbout's view, a scientific article is merely advertisement of scholarship; the real scholarship includes software and data which went into producing the article (a view taken by Pouzat as well, whose formulation I am paraphrasing [6]). Also at Stanford, RR has been pioneered by Buckheit and Donoho [7]. Greyer on his site lists a whole set of possible requirements for the research to be reproducible [4], such as "Anything in a scientific paper should be reproducible by the reader." That means results, plots, graphs. Again, how often do we do that? While these thoughts might sound discouraging, they should not be. It should be much easier for us to create RR than for life sciences; after all, we have it all in an electronic form (hopefully).

In our community, a recent opinion piece by Barni and Perez--Gonzales [8], spurred a few of us to start thinking seriously about the issue. In particular, Vetterli has been active in promoting the idea and his lab has already started thinking of ways of enabling RR [9]. We will mention those later in the paper.

1.4. Why Do We Need Reproducible Research?

Without sounding clichéd, the first reason should be to promote good science. We want science to be open and build upon previous work. Imagine if a mathematician had to start from scratch every time when trying to prove a new theorem. When approaching a new problem, we should be able to download all published algorithms pertaining to that problem and try them out, as well as be able to use the previous work as a building block for the future one. How many times has the DFT been implemented before the advent of Matlab (and even then)? It is to our advantage to use the collected knowledge and skip the tedium of developing something that already exists.

2. HOW DO WE GET TO REPRODUCIBLE RESEARCH?

While we all probably agree that RR is a good idea in principle, when it comes to the daily grind, lack of time, lack of space in a paper, problems start. However, if we all believed this is as essential to the paper as the proof and/or results are, then solutions can be found. Thus, in essence, the issue is our current culture. How do we change it to make RR an integral part and how do we encourage researchers to subscribe to the idea?

2.1. Issues to Consider

Cultural issues. Our culture prizes innovation above all else. In the *IEEE Transactions on Image Processing*, the reviewer is asked the following questions to rate the content of the paper: 1. Is the paper technically sound? 2. Is the coverage of the topic sufficiently comprehensive and balanced? 3. How would you describe the technical depth of the paper? 4. How would you rate the technical novelty of the paper? It is clear that questions 1. and 4. bear most weight in everyone's mind. Thus, novelty is of great importance to us. While there is nothing wrong with novelty, this criterion can lead to some strange situations. For example, in our "publish or perish" environment, one may see work which is novel but seems like an arbitrary exercise. Disconnected from the real world, we state our own problems, thereby fixing the assumptions so we can get something new (just think of the Gaussian assumption in anything and everything). On the other hand, the work which uses a known algorithm and then modifies it to suit a particular application, is typically considered lower-class. While a mathematician within us might think that a known algorithm that works in a particular application and on a particular data set is a sufficient proof of concept, we should examine its intrinsic value. A host of works developing a family of algorithms all based on the same "mother" algorithm, and which would work in a wide range of applications and on a wide range of data sets, would be most welcome, not to mention useful. We do not encourage such work, however.

Educational issues. When we come to educating our students, we do not do a good job of stressing the above values. Our students are typically undertrained in statistics and as a result might think that performing one experiment on one image should suffice. They typically reimplement everything without looking to find whether such pieces of code already exist. We have no set standards on how such code should be written or shared. While we might pay lip service to RR in principle, it is very hard to enforce those rules on a daily basis.

Data issues. Many of us work on data sets which we did not acquire. We might be collaborating with biologists, geophysicists, medical doctors, etc. The data given to us is someone else's hard work and our task is typically (though this is changing as well) to perform some type of signal processing. When the paper is prepared for publications, the issue of whether the data can be made available might not depend on us.

Intellectual property issues. Many in our community work with companies and various agencies which prohibit public disclosure of the code. In the same vein, in such situations, often very little detail is given on how the actual algorithm is developed. While this is a genuine issue, such work cannot be validated by others and should not have the same standing as the work which can be reproduced. One may decide to believe the authors, but since the work cannot be used by others, it does not benefit anyone except the company and the authors.

Collaborative issues. Our collaborations are varied: We collaborate with our students, colleagues from the same field at our universities or in our companies, colleagues from different universities, colleagues from other fields. In each of these instances, a variety of problems might arise, some of which, such as those pertaining to data and intellectual property, have been discussed above.

2.2. Suggested Course of Action

The above thoughts suggest the following:

- Encourage authors to publish first-class, experimental work.
- Encourage authors to submit work which uses a known algorithm in a new setting or with a different type of data.
- Show value of such work by publishing special issues, promoting it through paper awards and training students to perform such work.
- Have a blueprint of what should be done once the paper is accepted for publication. The authors should have code producing

the results, all tables and figures should have accompanying code replicating them, a readme file should be included to explain the usage and data used in the experiments should be made available if possible. See next subsection on details.

• In cases where data is not ours and issues with allowing public access to the data exist, negotiate for a representative sample to be available.

• Promote the idea of RR with the national funding agencies. Develop templates of what should be published and how. Develop templates for collaborative work and sharing of data.

2.3. How Do We Publish Reproducible Research?

It is unlikely that anything can be done overnight. While we might strive for RR as our guiding principle, it will be years before it becomes standard practice. In the meantime, what can we do to encourage and reward "good behavior" (All of you parents out there understand what I am talking about.)?

One idea floating around is to have our Transactions have a special section with a heading "RR Papers". These papers will thus be prominently displayed and will consequently carry "more weight". Another is to possibly establish a paper award for an RR paper thus sending a message this is something we take seriously. We should strive to form a rough guideline of what each paper should contain and what should be the accompanying material to make it worthy of the "RR" designation. Such papers would earn the right to prominently display "RR" on the first page. We should strive to include comparisons to appropriate RR works in our own papers.

How to Write RR Papers. We should establish good practices of how we write our papers. The following ideas with some modifications have been extracted from a document prepared by Mauro Barni to spur the discussion on RR. A paper being published as RR should satisfy some or all of the following:

• A block diagram together with the pseudo-code should be included as well as the description detailed enough to allow readers to reimplement the algorithm with no uncertainty.

• All the parameters that are needed to run the algorithm should be clearly listed in a table and the values used in the experiments reported in the paper.

• The URL with the software implementing the proposed algorithms available to both reviewers and/or the readers.

• The information about the data used to run the experiments is clearly defined or made available to the readers/reviewers when possible. Gentleman and Lang in [10] call the above a *compendium*: "a container for the different elements that make up the document and its computations (i.e. text, code, data, ...), and as a means for distributing, managing and updating the collection." To enable this point, the creation of suitable public databases should be encouraged.

How to Make Papers RR. On the LCAV site [9], suggestions are made on how to make the paper reproducible. This includes the following instructions:

"Make a web page containing the following information:

1. Title.
2. Authors (with links to the authors' websites).
3. Abstract.
4. Full reference of your paper, with current publication status, and a PDF of your paper.
5. All the code to reproduce all the results, images and tables. Make sure all the code is well documented, and that there is a readme file explaining how to execute it.
6. All the data (images, measurements, etc) to reproduce all the results, images and tables. Add a readme file explaining what the data represent.
7. A list of configurations on which you tested your code (software version, platform).
8. An e-mail address that people can use for comments and remarks (and to report bugs).

Depending on the field in which you work, it can also be interesting to add the following (optional) information to the web page:

1. Images (add their captions, so that people know what Figure xx is about).
2. References (with abstracts)."

While initially, we might individually decide on how we want to approach RR, hopefully, there will be movement throughout the Society to standardize these.

Tools to Enable RR. In the past decade, tools have emerged to help enable RR. These tools offer environments for reproducible computational experiments and greatly simplify maintenance of software. Sweave offers literate programming for RR [11, 4]. In [2], the authors develop SCons, open-source tools designed for building software for reproducible computational experiments. Obvious issues with software and data include where the repository should be (on the author's site or on the journal's site) as well as the software becoming obsolete and data not being available. These are issues that have to be discussed in a larger forum and perhaps this is an opportunity to start such a discussion.

3. AN ENTIRELY NON-RR CASE STUDY

Having discussed a number of issues and a number of things we could do, I will now proceed to present a short, entirely non-RR case study. The data set I used consists of 15 papers published in the past few years in the *IEEE Transactions on Image Processing*. I chose an EDICS category that is both theoretical and experimental and have made sure all of the 15 papers both propose new theoretical models/tools and then build new algorithms based on those. I stayed away from standard-oriented EDICS categories as well as the newly introduced biomedical ones (though initially I wanted to compare against those and see if different trends emerge, I left this for future work). For all of these papers, competing algorithms exist. For some of the application fields chosen, public databases exist. I then read those 15 papers and rated them on a scale of (0, 0.5, 1) on the criteria I divided into two sets:

• *Algorithm and Experimental Setup:* In this part, I rated papers on (a) how well they explained the algorithm details, (b) how well the data used was explained, (c) the data size, (d) details on parameters used and (e) comparison to competing algorithms. For each of these, the paper got 0 if it failed the criterion completely, 1 if the criterion was completely satisfied and 0.5 if it fell somewhere in between. Just remember this was a purely subjective exercise.

• *Reproducible Research:* In this part, I rated papers on (a) whether they had a block-diagram of the algorithm (b) whether they had pseudo code of the algorithm, (c) whether the data was available, (d) whether the code was available and (e) whether the proof was available. In (c), if the authors used a public database and identified it earned the paper a 1. In (d), I searched both in papers as well as authors' websites to see whether the code for the algorithm was available anywhere.

The results of my investigation are given in Table 1. Looking at these numbers, we note several interesting facts: (a) All papers

Algorithm and Experimental Setup [%]					Reproducible Research Criteria [%]				
Algorithm details	Data details	Data size	Parameter details	Comparisons	Block diagram	Pseudo code	Data available	Code available	Proof available
80	33	46	46	26	0	60	33	0	100

Table 1. Case study: Data collected for 15 papers published in the *IEEE Transactions on Image Processing* in the past few years. The ratings were performed by the author on the scale of (0, 0.5, 1) with 0 being unsatisfactory and 1 being satisfactory.

had proofs, while none had code available. (b) The algorithms were typically explained in sufficient detail but none of the papers had a block-diagram of the algorithm. Given that the block-diagram is usually the easiest way to visualize an algorithm, this is fairly strange. (c) The facts related to the details on data used, data size and availability of the data are all below average. In all of the cases where I rated data availability as 1, the authors identified a publicly available database. (d) Only in about half the cases were the parameters specified. Actually, the whole set of parameters was given in fewer than half the cases and those I rated 1. There were few I rated 0.5 for specifying at least some of the parameters. (e) In only about a quarter of the papers did the authors actually compare the algorithm against competing algorithms. I believe this to be the result of (b)-(d); one cannot replicate someone else's algorithm if no sufficient detail nor parameters are specified in a satisfactory way. (f) I was pleasantly surprised to find that in 60% of the cases, pseudo-code was available.

Having said all this, how would I rate myself on the above criteria. I would say 0 on algorithm details, data details and parameters. There is really no competing algorithm so this would be a Not Applicable (NA). I did not give you a block-diagram, pseudo code, will not make this data available, there is really no code and the proof is an NA. So you are left to believe me when I give you the above numbers. Should you? Of course not, unless you can recreate these numbers yourself. (If you have time, you can entertain yourself with the same exercise and see what you come up with).

4. CAN WE MAKE IT HAPPEN?

While the above account abounds with problems and obstacles, I believe we are all scientists because we want to make a mark. The best way for our work to be recognized and make a true difference is for it to be shared and used by as many people as possible. RR is an obvious route towards this end, and thus, I believe we will get there; it is in our best interest.

5. ACKNOWLEDGMENTS AND DISCLAIMERS

I would like to thank Mauro Barni and Fernando Perez-Gonzalez for starting the whole discussion in our community by publishing their column in the IEEE Signal Processing Magazine [8]. I believe those two pages will have a huge impact on how we do our research. My gratitude goes to Martin Vetterli who pointed the column to me, involved me in the discussions, started the ball rolling and most of all, for his never-ending enthusiasm for thinking outside the box and doing things the right way. His group has taken concrete steps in the RR direction [9]. Patrick Vandewalle deserves our thanks for organizing the special session at ICASSP 2007. I thank the members of the *IEEE Transactions on Image Processing* Editorial Board for their thoughtful examination of issues pertaining to publishing RR papers. Finally, thanks to the informal email group Mauro Barni and Fernando Perez-Gonzales organized for sharing their opinions and suggestions.

The thoughts expressed here, when not cited, are usually my opinions on the issue. As such, these are necessarily my personal views, colored by five years I spent as the Editor-in-Chief of the *IEEE Transactions on Image Processing*. All mistakes and omissions fall squarely on my shoulders. I welcome your input and suggestions in hope that in the near future, we will be closer to the our ideal of open and reproducible science.

6. REFERENCES

[1] R. J. LeVeque, "Wave propagation software, computational science, and reproducible research," in *Proc. Int. Congr. of Mathematicians*, 2006.

[2] S. Fomel and G. Hennenfent, "Reproducible research philosophy," http://egl.beg.utexas.edu/RSF/book/rsf/scons/-paper_html/node2.html.

[3] D. E. Knuth, "Literate programming," *Computer Journ.*, vol. 27, pp. 97–111, 1984.

[4] C. Greyer, "Sweave demo," http://www.stat.umn.edu/∼charlie/Sweave/.

[5] "Stanford Exploration Project," http://sepwww.stanford.edu/.

[6] C. Pouzat, "Reproducible neurophysiological data analysis?," http://www.biomedicale.univ-paris5.fr/SpikeOMatic/-Compendium.html.

[7] J. Buckheit and D. L. Donoho, *Wavelets and Statistics*, vol. 103, chapter Wavelab and reproducible research, pp. 55–81, Springer-Verlag, 1995.

[8] M. Barni and F. Perez-Gonzales, "Pushing science into signal processing," *IEEE Signal Proc. Mag.*, Jul. 2005.

[9] "Reproducible research page of the LCAV Lab at EPFL, Lausanne," http://lcavwww.epfl.ch/reproducible_research/.

[10] R. Gentleman and D. T. Lang, "Statistical analyses and reproducible research," *Bioconductor Project Working Papers*, , no. 2, May 2004, http://www.bepress.com/bioconductor/paper2.

[11] F. Leisch, "Sweave," http://www.ci.tuwien.ac.at/∼leisch/.