Eurographics Workshop on 3D Object Retrieval (2015)
I. Pratikakis, M. Spagnuolo, T. Theoharis, L. Van Gool, and R. Veltkamp (Editors)

# Towards Scientific Benchmarks: On Increasing the Credibility of Benchmarks

Odd Erik Gundersen

Department of Information and Computer Science
Norwegian University of Science and Technology
Trondheim, Norway

**Abstract**

***Problem:*** *Increasing the credibility of results from scientific benchmarks.* ***Goal:*** *Specify what exactly is required in order for a benchmark to be scientific.* ***Contribution:*** *(i) Specification of what it entails for a benchmark to be scientific, (ii) a metric for measuring the replicability of an experiment, (iii) a metric for measuring the replicability of a set of experiments, and (iv) Analysis of the replicability of SHREC 2015.* ***Result:*** *Replicability of SHREC 2015 can be increased by open sourcing the methods compared and improving documentation.*

Categories and Subject Descriptors (according to ACM CCS): I.5.2 [Pattern Recognition]: Design Methodology—

## 1. Introduction

The problem we investigate in this paper is related to replication of scientific benchmarks, and what is required in order to trust their results. More specifically, we analyze what a scientific benchmark is and we apply the outcome of this analysis to 3D object retrieval benchmarks in order to shed light on how to produce the most trustworthy results in such benchmarks. So, what exactly is a scientific benchmark? According to the Merriam Webster Dictionary[†], a benchmark is defined as *"a standardized problem or test that serves as a basis for evaluation or comparison (as of computer system performance)"*. Scientific is defined by Merriam Webster Dictionary as *"done in an organized way that agrees with the methods and principles of science,"*, and hence science must be defined. According to the American Physical Society (NPS) *"Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories. The success and credibility of science are anchored in the willingness of scientists to: 1) Expose their ideas and results to independent testing and replication by others. This requires the open exchange of data, procedures and materials. 2) Abandon or modify previously accepted conclusions when confronted with more complete or reliable experimental or observational evidence. Adherence to these principles provides a mechanism for self-correction that is the foundation of the credibility of science.*[‡]*"* We see that replication is one of the foundations of science.

Before continuing the discussion, we need to distinguish between replication and reproducibility, which often cooccur in discussions about the principles of science, but have their own distinct meanings. Replication differs from reproducibility in being an exact duplication of an experiment. If the result of an exactly duplicated experiment is the same as the results from the original experiment, then the results are replicable, and thus the experiment is valid in a scientific sense. Reproducibility, on the other hand, is causing the same result in an experiment that is similar, but not an exact copy. A similar experiment can be conducted by, for example, performing the exact same experiment on a different data set or carrying out a computer science experiment using a reimplemented version of the method under investigation using a different programming language. As replication is an exact duplication of an experiment, it requires an open exchange of data, procedures and materials (research mate-

[†] URL: http://www.merriam-webster.com/, accessed 14th of February 2015.

[‡] URL: http://www.aps.org/policy/statements/99_6.cfm, accessed 14th of February 2015.

DOI: 10.2312/3dor.20151059

rials as defined by Tufts University[§]). Hence, the results of a scientific benchmark can be trusted when all data, procedures and materials are exchanged.

Now, what does this mean for 3D object retrieval benchmarks? First of all, 3D object retrieval can be interpreted as a learning task [Mit97], which means that parts of the data are used for training the learner and parts are used for testing. Typically, the data set is divided into three parts: One data set is used to train the learner (training set), one set is used to adjust parameters or select models (validation set), and last data set is used to assess the retrieval error (test). In addition to specifying which data sets have been used for benchmarking, a specification of which data samples that belongs to the different data sets must be provided in order for someone to replicate a benchmark. Furthermore, as 3D object retrieval benchmarks is a branch of computer science, procedures are encoded and represented as software. The software that is required for carrying out the experiment (benchmark) includes the methods that are benchmarked, but also the software that configures the experiments. The experiment configuration includes specifying the parameters of the methods that are used when running them, such as specifying the number of neighbors used in the k-nearest neighbor algorithm. The material includes papers and presentations that describe the methods, pseudo code as well as documentation. It is not enough to provide the source code, the source code has to be well documented and transparent, so that understanding it becomes as simple as possible. Other material includes specifications of the hardware and software used (including version) for running the methods that are benchmarked. Differing infrastructure, such as operating systems and hardware, will influence the results, especially if efficiency (how fast a method executes) is benchmarked. However, running benchmarks on different infrastructure might also impact the performance (how accurate a method is) because of the problems inherent to floating point calculations [HKJ*13].

**Goal:** The goal of the research presented here is to specify what exactly is required in order for a benchmark to be scientific. A benchmark should be scientific, as this will be increase the credibility of the benchmark. We will apply our specification to SHREC benchmarks to see whether they are scientific according to the specification.

**Contribution:** The contribution is fourfold: (i) what it entails for a benchmark to be scientific is defined, (ii) a metric for measuring the replicability of an experiment is proposed, (iii) a metric for measuring the replicability of a set of experiments is proposed, and (iv) the replicability of SHREC 2015 is analyzed using the proposed metrics.

## 2. The Challenges of Computer Science

According to [Ioa05] 85% of all research findings are false. There are several reasons to this, including bias in the experiment design and lack of testing by more than one team. Artificial intelligence research is mentioned as one of the worst domains because of flexibility of designs, definitions, outcomes, and lacking analytical modes. Benchmarks, such as the ones in SHREC 2015, where predefined and common evaluation criteria are utilized counteract this negative trend.

Many fields of science have problems with reproducing published research findings that have been published in high impact journals. Replication of many psychology experiments is seemingly impossible as the experiments are performed on human beings, and the effect typically can wear off if the same population is exposed to the same experiment twice [Yon12]. One would think that replication is simpler for computer science. However, this is not the case, which is increasingly being realized by the computer science field. Several efforts have been started, and in 2009 a special issue of Computing in Science & Engineering focused on reproducible research [FC09]. Making computer science experiments fully executable in such a way that documentation and figures have been automatically generated have been suggested by several research groups [KdE11] and [BD95].

The three main requirements for producing reproducible machine learning research are (i) open source software, (ii) open data and (iii) open access papers according to [BO14]. Workshops have been organized [LMS12], manifestos defining and arguing for replication/reproducibility of computer science experiments have been proposed [GEN13]. However, challenges still exist [NAB*12]. One of the research practices that is suggested for increasing the proportion of true research findings is reproducibility practices [Ioa14].

## 3. Requirements of a Scientific Benchmark

Ideas and results should be exposed for independent testing and replication in order to be scientific. Thus, transparency is required into the both the reasoning and the experiments that have been conducted, which is specified as an open exchange of data, procedures and materials in the definition. Full disclosure of everything related to the scientific endeavor provides confidence in the scientific work and its results. With full disclosure comes a high level of credibility. However, full disclosure alone does not provide the highest possible level of credibility to a scientific work. The highest possible level of credibility can only be given to scientific work that is replicated over and over again by other research groups in different labs than the research group performing the research in the first place. Only when such third party replications have been conducted and properly documented can the scientific work be considered fully credible. This is a high standard to measure science against, and such a standard requires scientists to document their research meticulously, as well as providing all materials.

[§] URL: http://sites.tufts.edu/dca/collections/collection-policies/collection-policy-for-faculty-papers/research-materials/, accessed 15th of February 2015.

Let us have a look at what full disclosure and replication of a computer science experiment that has been conducted fully on a computer means. We restrict this discussion to experiments performed fully on a computer, such as simulations, learning tasks, computer performance measurements and similar experiments, and we use computer experiments to designate such experiments in this paper. Hence, the discussion leaves out human computer interaction experiments where users are part of the experiment. The requirements for full disclosure can be summed up by:

**What** has been done is specified by the software program (executable) running the experiment, the data (if any) that has been used and the results.

**How** it is done document the methods encoded by the executable program, as well as the evaluation criteria. The executable program includes the source code that the executable has been compiled from, software documentation and pseudo code.

**Why** documents the motivation for conducting the experiment in the first place, and it includes notes, presentations, technical reports and scientific papers.

**Where** describes the environment and infrastructure of the experiments and includes the hardware specifications and specification of the software used for running the experiments (including version), such as the operating system or scientific software such as Matlab, R, Mathematica to name a few.

**Who** describes who has done it and thus whether it is novel work or a replication.

A benchmark is a standardized test or problem that serves as a basis for evaluation or comparison. This means that some of the items in the list of requirements for full disclosure are restricted, such as what, where and who. Performance benchmarks, such as GPU performance, is typically done by a third party (who) tests some software performing one or more specified tasks (what) in a given environment such as hardware and software platform (where). A scientific benchmark should be fully disclosed in a similar fashion to any other scientific experiment, but in addition it should be disclosed on beforehand, so that those involved in the programs being benchmarks can contribute to iron out any misunderstandings.

## 4. Measuring the Replicatibility of a Scientific Body of Work

The above list of requirements can be broken down along the three dimensions (i) experiment procedures, (ii) data and results, and (iii) documentation, which can be used as a basis for developing a metric for measuring the replicability of a scientific body of work. Let us have a closer look at them: **Experiment Procedures**. The actual procedures representing a computational experiment is contained by the software and its source code: (i) source code - method, (i) source code - experiment Setup, (iii) executables. **Data and Results**. The data that has been used to train and test the method as well as the results it produces: (i) Training data, the actual samples used for training, (ii) validation data, the actual data samples used for validation, (iii) test data, the actual samples used for testing, (iv) results, the actual results produced by the experiment, and not only a graph representation of it. **Documentation**. All material that documents the method and the experiments: (i) Experiment documentation including motivation for using a given data set and the chosen parameters. (ii) Textual explanation of the method, (iii) pseudo code, (iv) software documentation including the code documentation, (v) infrastructure (hardware and software, including version information), (vi) evaluation criteria, such as evaluation metrics.

We can express a simple, metric function, $R$, that measures the replicability of an experiment, $e$, in the following way:

$$R(e) = \frac{1}{3}[Proc(e) + Data(e) + Doc(e)], \tag{1}$$

where $Proc(e)$, $Data(e)$ and $Doc(e)$ are metrics for measuring how well the procedures can be replicated, how much of the data and results are available and how well the experiment is documented, respectively. The output range of these four functions is [0.0, .., 1.0]. A lower score indicates worse replicability and 1.0 indicates a fully replicable experiment. $R(e)$ can be expressed using weighted averages, and the weights can be set according to what is interpreted as more or less important for replication. However, for simplicity and increased readability, in addition to problems with identifying the weights, we will omit weights. $Proc(e)$, $Data(e)$ and $Doc(e)$ are three functions defined as follows:

$$Proc(e) = \frac{1}{3}[SCM(e) + SCE(e) + EXEC(e)], \tag{2}$$

where $SCM(e)$, $SCE(e)$ and $Exec(e)$ are boolean functions that are true if the source code is open, the source code for the experiment setup is open, or the executables are open, respectively.

$$Data(e) = \frac{1}{4}[TRN(e) + TST(e) + VAL(e) + RES(e)], \tag{3}$$

where $TRN(e)$, $VAL(e)$, $TST(e)$ and $RES(e)$ are boolean functions that are true if the training data is open, the validation data is open, the validation data is open, or the results produced by the executable is open, respectively

$$Doc(e) = \frac{1}{6}[EXP(e) + MET(e) + PC(e) + CD(e) + INF(e) + EC(e)], \tag{4}$$

where $EXP(e)$, $MET(e)$, $PC(e)$, $CD(e)$, $INF(e)$ and $EC(e)$ are boolean functions that are true if an textual description of the experiment is open, an textual description of the method is open, the pseudo code for the method is open, the source code for the software describing the method and the configuration is properly documented and open, the infrastructure (hardware and software) used is documented and open, or the evaluation criteria is documented and open, respectively. Based on $R(e)$, we suggest a metric that computes the replicability of a set of $n$ experiments where $e_i$ represent individual experiments in this set:

$$Replicability(E) = \frac{1}{n}\sum_{i=1}^{n} R(e_i). \quad (5)$$

## 5. Experiments: Measuring Replicability

In this section, we measure the replicability of a paper, a full conference track and SHREC 2015. We use the complete first track from the proceedings of IJCAI, the Agent-Based and Multiagent Systems track, to calculate the replicability of a conference track and the first paper of this track to calculate the individual paper. The information needed is collected as part of a structured literature review. However, information about the following requirements were not collected: executables, $EXEC(e)$, result outcomes, $RES(e)$, pseudo code, $PS(e)$, source code documentation $CD(e)$ and evaluation criteria, $EC(e)$. $EXP(e)$ is evaluated as true if experiment parameters are discussed. $MET(e)$ is evaluated as 1.0 for all the scientific papers, as the information is gathered from scientific papers, and hence it is omitted from the calculation. For these examples, infrastructure is evaluated accordingly: $INF(e) = \frac{1}{2}[SW(e) + HW(e)]$. The data used in this example can be found at the authors website¶.

**Paper:** None of the data were shared, so the $DATA(e)$ score is 0.0. Furthermore, none of the source code that encoded the experiment were shared, and hence the $PROC(e)$ score is 0.0. However, both the hardware setup was specified as well as the experiment parameters, but not the software environment, and thus the documentation score calculated using $DOC(e)$ is 0.67. Thus, $R(e) = \frac{1}{3}[DATA(e) + PROC(e) + DOC(e)] = 0.22$.

**Track:** 58 scientific papers were published in the track. Out of them 20 where theoretical and did not document experiments, and hence, $n = 38$. As only 3 out of the 38 papers use open data and none of these three document which data that has been used for validation and test, the combined $DATA(e)$ for all the papers in $E$ score is very low at 0.03. Similarly, the combined score for all papers for the experiment procedures $PROC(e)$ is low (0.013), as only one paper refer to open source software. However, only the method was shared and not the experiment setup. Although a much better score on the combined documentation $DOC(e)$ for all the papers, the total is not impressive at 0.40. In total $Replicability(e) = 0.15$ for a track on one of the most prestigious AI conferences.

**SHREC 2015:** Each track benchmark different 3D object retrieval methods using open data sets with differing characteristics. Both training and test data is provided ($TRN(e) = 1$ and $TST(e) = 1$), and the evaluation methods ($EC(e) = 1$) are described as part of the benchmark description. It also contains a description of the experiment that is to be performed ($EXP(e) = 1$). The result of each track is a paper summarizing the retrieval methods ($MET(e) = 1$) and the results of applying them to the open data sets ($RES(e) = 1$). This paper is not required to contain pseudo code of the methods. The source code is not required to be open (one track require it for a review by the organizers), and three tracks require the executables. Hardware infrastructure is only required by one of the workshops as efficiency is a part of the evaluation. Thus, $Replicability(SHREC2015) = 0.6$. In order to increase the replicability of SHREC benchmarks, and by this making them more scientific, the participants should be encouraged to open source the methods. Some adjustments can be made to the documentation of the benchmarked methods as well.

¶ Data: www.idi.ntnu.no/~odderik/3DOR/

## References

[BD95] Buckheit J. B., Donoho D. L.: *Wavelab and reproducible research*. Springer, 1995. 2

[BO14] Braun M. L., Ong C. S.: Open science in machine learning. *Implementing Reproducible Research* (2014), 343. 2

[FC09] Fomel S., Claerbout J. F.: Reproducible research. *Computing in Science & Engineering 11*, 1 (2009), 5–7. 2

[GEN13] GENT I. P.: The recomputation manifesto, 2013. 2

[HKJ*13] Hong S.-Y., Koo M.-S., Jang J., Esther Kim J.-E., Park H., Joh M.-S., Kang J.-H., Oh T.-J.: An evaluation of the software system dependency of a global atmospheric model. *Monthly Weather Review 141*, 11 (2013), 4165–4172. 2

[Ioa05] Ioannidis J. P. A.: Why most published research findings are false. *PLoS Med 2*, 8 (08 2005), e124. 2

[Ioa14] Ioannidis J. P.: How to make more published research true. *PLoS medicine 11*, 10 (2014), e1001747. 2

[KdE11] Kauppinen T., de Espindola G. M.: Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science 4* (2011), 726–731. 2

[LMS12] LeVeque R. J., Mitchell I. M., Stodden V.: Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science and Engineering 14*, 4 (2012), 13. 2

[Mit97] Mitchell T. M.: Machine learning. 1997. *Burr Ridge, IL: McGraw Hill 45* (1997). 2

[NAB*12] Neylon C., Aerts J., Brown C., Coles S. J., Hatton L., Lemire D., Millman K., Murray-Rust P., Perez F., et al.: Changing computational research. the challenges ahead. 2

[Yon12] Yong E.: In the wake of high profile controversies, psychologists are facing up to problems with replication. *Nature 483* (2012), 298–300. 2