Using Computer Simulations for Producing Scientific Results: Are We There Yet?

Olivier Dalle
University of Nice Sophia Antipolis & INRIA

March 2013
What are Scientific Results?

• For the Scientist that produces them: the (cross-)product of his labour and expert knowledge
  – Phew... I worked hard on this! This is a good one!

• For whom pays for it: Publications!
  – The metric is easy... Even though it is scientifically arguable!

• For many: very complicated stuff that nobody can understand
  – Including scientists in other disciplines
What are Scientific Results?

- Obvious answer:
  The by-product of a **scientific** approach
- So... What is a scientific approach?
- What is science?
  - Giving a syntaxique definition is almost easy
    - knowledge about things based on observations, experiments, theories and demonstrations
    - Exact science, experimental science, human science
  - Explaining how to do good Science
    - A long standing philosophical question...
What is Good Science?

Aristotle
384 – 322 BC

Many scientific concepts
- The nature of truth
- Logic & Reasoning
- Demonstration
- Dialectics
- Theory of universal
- Metaphysics/Abstractions
- 4 types of causes
  - Material
  - Formal
  - Efficient
  - Final
What is Good Science?

- Middle-age... Dark times...

Thomas Aquinas
1225 – 1274

Scholastics:
Follow exclusively established principles & beliefs without questioning...
Dangerous times!
What is Good Science?

- Age of reason. Enlightenment

Started with Copernic...
(1473 – 1543)

... Followed by Galileo...
(1564 – 1642)
- Combines Experiments and Maths
- Laws of nature are mathematical

...And Descartes.
"Early" Modern Doctrines

- Cartesianism (Descartes, 1596-1650)
  - Cartesian doubt
- Rationalism: reasoning comes first
  - Discourse on Method (Descartes)
  - Determinism (Spinoza)
  - Leibniz' monads (units of reality)
- Empiricism: observing comes first
  - Actually not so modern (since Aristotle), but went though ages
What is Good Science?

Modern Doctrines

• (American) Pragmatism
  – Instrumentalism (eg. Milton Friedman)
    • Scientific theory is an instrument to understand the world
      – A black-box that produces predictable outputs in response to inputs...
    • Does not deny existence of postulated entities
    • Kuhn's Structure of Scientific Revolutions
  – Also includes radical empiricism, verificationism, conceptual relativity, fallibilism...
• Logical positivism, Logical empiricism, neopositivism...
• Simulism!

Olivier Dalle – Using Simulation for Producing Science
What is Good Science?

Tentative answer:

A. Problem solving
   - joining instrumentalism's black-box idea

B. Contributes/extends understanding/knowledge

C. A Communication exercise/problem
   - Claim: Secret/unknown work is not science
   - Science need to discussed, challenged, built upon,...
   - Actually used for assessment [with questionable metrics]

D. No universal approach
   - Although some widely accepted "idioms"
     • There must be room for novelty [Kuhn's SSR]
   - Independence of reasoning [avoid scholastics]
Talk Outline

Let's go for a bit of dialectics!

- Hypothesis: Something's wrong with reproducibility
- Thesis: Reproducible Research
- Anti-thesis: Drummonds' paper
- Synthesis: Some Ideas & Solutions
Hypothesis:
SOMETHING'S WRONG WITH REPRODUCIBILITY
A Rising Concern...

- Schwab, Karrenbach, and Claerbout (2000)
  - Reproducible research applied to computation
- Pawlikowsky et al. (2002)
  - Survey 2200 publications on telecom networks: majority do not satisfy criteria of credibility
- Perrone, Yeong, and Ruth Lee (2003)
  - Hidden details end up missing in publications
- Kurkowsky, Camp, and Colagrosso (2005)
  - Survey of MANET studied: 75% use simulation, but less than 15% are repeatable
Now an Established Issue

• Ince, Hatton, and Graham Cuming 2012
  – Editorial of *Nature*
• *Biostatistics Journal*: Associate Editor for Reproducibility
• *Science* Mandates source code
  – A first step, but maybe not sufficient
• Olivier Dalle (and others I guess!), 1999-present
  – Almost Never able to reproduce simulations from others' papers
  – Can't work to extend or validate published results without redoing everything from scratch (re-solving the problem)
Thesis:
REPRODUCIBLE RESEARCH (RR)
Reproducible Research

Name of a movement/group

• Based on Jon Claerbout's views in 2000
• Multi-disciplinary
• Workshops
  Eg. Reproducibility in Computational and Experimental Mathematics (ICERM, 2012), AAAS Annual meeting, Workshop on the digitization of science (2011), Applied mathematics perspectives workshop on reproducible research (2011), ...
• Papers...
Definitions

ICERM report (Stodden et al., 2012):

- “reproducible research” and “reproducibility” most often refer to the ability to recreate computational results from the data and code used by the original researcher.
- "related to but distinct from [...]":
  - "numerical reproducibility" (same program gives different results)
  - "repeatability" (an experiment is conducted independently from first principles)
Motivations

Stodden 2010 (also cited by Drummond, see later)

- "Generating verifiable knowledge has long been scientific discovery’s goal"
- Adhere to scientific method
  - not clearly defined, but not easy
- Build on previous generations achievements
- Avoid controversies and retractions
- Tradition of published science. A paper presents:
  - New results
  - Material needed for reproducibility
Motivations

Later, from ICERM report (Stodden et al., 2012)

- Verify findings
- Efficiently build on past results
- Apply basic tenets of Scientific Method to computational procedures
- Align on "bench scientists" practices
  - [They] have Lab notebooks, extensive documentation
  - [We] miss workflows and code
- Productivity, Credibility, Community
Compendium
(Material needed for RR)


a. Research paper
   1. Including sources
Compendium
(Material needed for RR)


a. Research paper

b. Data

1. Data
2. Documentation [source, components, interpretation]
3. How data was produced
4. Code
5. Code documentation
Compendium
(Material needed for RR)


a. Research paper
b. Data
c. Experiment
   1. Code and instructions
   2. Code documentation, algorithms
   3. Listing of parameters
   4. Experimental methodology
Compendium
(Material needed for RR)


a. Research paper
b. Data
c. Experiment
d. Results

1. Figures
2. Documentation and explanation
3. Auxiliary material
4. Code used for presentation
5. Doc of auxiliary code
6. Description of computing platform
Anti-Thesis:

DRUMMONDS' PAPER
A dissenting Opinion

Restates RR arguments as follows:

• "RR is an essential part of science, not doing so is bad science"
• "Important step of scientific method to build on previous work, or progress slows"
• "Requires submission of data and tools used to generate results, or it cannot be verified and built upon"
• "RR necessary to prevent misconduct, or crisis of confidence in science"
1. Essential part of science?
   - There are quite different views on what replicability means
   - Two ideas, three concepts:
     - Reproducibility
       - duplicate experiment as far as possible, to minimize difference with original work
     - Statistical replicability
       - Avoid results due to chance due to limited sample
     - Scientific replicability
       - Robustenss and generalizability: increase difference to measure change
   - Statements:
     - It seems clear to me that reproducibility as proposed by the round-table has never been a central tenet of science.
     - Only Scientific Replicability has any real claim to be a gold standard.
2. Important step of scientific method?

   - The idea of a single scientific method is pervasive and to many persuasive:
     - Taught at every level of education
     - Identify what falls under the rubric of science
     - Simple steps that if followed should produce solid science
   
   - Statement: See modern philosophy (e.g., Kuhn's SSR)
     - Claim of single scientific method is debatable
     - Requiring particular steps is suspect
Point-by-point discussion

3. Require submission of data and code?
   - Publishing code is almost free
   - Statements
     • No wish to argue against voluntary submission
     • Submitting code will result in accumulation of questionable software
     • Waste of time since most software will not be used
       - Many papers are uncited
     • Questions the idea of "system with devoted scientific community that maintains code and reproducibility"
       - Lots of efforts, undermine levels of trust between researchers
       - Policing would have negative consequences
4. Solve misconduct issues in Science?
   - RR states a recent increase
     • Credibility crisis
   - Statement: Misconduct is not new in Science
     • Some published results of Mendel are considered too good to be true
     • Newton is not above reproach
     • The Case of "Cold Fusion"
       - Many scientists attempted to reproduce and failed
     • It is not the main reason why public has little trust
       - Science does not produce crystal-clear responses
Synthesis:
SOMES IDEAS
So, is RR good or bad??

Let's be pragmatic:

• I can't reproduce others experiments
  – I wish that concern would be addressed
  – I don't want this to come at a prohibitive cost

• I DO need it!
  – Not only to because I am suspicious, but mostly because I am curious!

• Compared to other Sciences
  – computer-based Science is becoming badly perceived
  – I want to be trusted and respected in my work
So, is RR good or bad??

My claims

• we DO need smart tools to support needs and methods
  - The burden should be on simulator/tools designers rather than users
• we must be pragmatic
  - Do not try to force people doing something counter-productive
  - Do not argue a priori about "good" science, ...
  - Come with useful ideas
More pragmatics

- Pre-formatted workflows are prone to fail meeting expectations
  - We need tools to support Science, not to frame it
- Most of RR efforts should still be paid by consumer, not producer
  - Assess cost/benefits for producers and consumers of Science
Examples

• Can we find an equivalent to lab notebooks?
  – Workflows?
    • Good way: keep record of experimental steps
      – Useful to archive experiments, not much additional work
      – Useful to replay steps of computerized study
    • Bad way: force users to follow a predefined workflow
      – Might not be adequate => effort to fit in or create a new one
      – Limit independence: there is no single scientific method

• Submit code along with publication?
  – A snapshot of source hierarchy on GitHub/SourceForge/GCode is easy for producer
  – A virtual machine is better but more tricky...
Simulation Distinctive Reproducibility Features

• Ability to Replay
  – Re-run the exact same simulation experiment with different settings, eg. different observation
  – may produce different outputs for the "same" simulation

• Fully Virtual Nature
  – Compared to "Un-virtual" sciences, subject of study is also virtual
  – Introduces a risk of reproducing study of wrong subject
Issues for Reproducibility

• Human Factors
  - Genuine ignorance (e.g. use a bad RNG)
  - Insufficiently detailed publication
  - Business related limitations
  - Manipulation errors

• Technical Issues
  - Software bugs (may induce unrepeatable results once fixed)
  - Software availability
  - Numerical computations (e.g. Floating Point arith. ...)
  - Computer platforms and OS evolutions
Solutions for Reproducibility

- Automation: limits the risk of human errors
  - Eg. SAFE (Perrone et al.),
  - AKAROA (Pawlikowsky et al.),

- Scientific workflows
  - Eg Taverna
  - Integrated solutions with James II, CD++, ...
  - Actually not so level independent
    - requirement also applies to workflow engines
Solutions for Reproducibility

• Ensure long-term execution of a simulation
  – platform independent solutions
    • eg. language, formalisms with strong semantics (like DEVS)
  – platform generation solutions
    • eg. kameleon.imag.fr
  – ensure availability of platform
    • eg. virtualization
Solutions for Reproducibility

• Archiving
  – Requirement: Retrieve Model & Scenario
  – Need to archive critical informations (Compendium)
    • Suggestion: this should be the role of publications
• Possible solution: rely on publishers & editors
  – Require submission of "source" code with publication
    • Enforced by editors & publishers (eg. Science)
  – Additional material becomes part of the publication
    • Bonus: Code can be cited
  – Review the additional material
    • Check that it allows for reproducibility
Pros & Cons of a Publication-Based policy

- **Cons**
  - Publishers need to adapt and support additional costs
  - Editors must include material in decisions
  - Peer-review process more complex

- **Pros**
  - Source code can be cited
    - Better recognition of scientific SW production
  - Allows reproduction and **extension** of previous works
  - Enforce use of Open Source software
    - But cons. if conflict with business model
Are we there yet?

Maybe not so far

- No "integrated" publication process
  - But parts: free software repos (GitHub, SourceForge, Google)
  - Some legal matters have to be clarified (software isbn? doi?)
- Integrated solutions in simulators
  - NS3: SAFE, ...
  - SimGrid: ANR SONGS Project, WP8: OpenScience
  - Model-DBs in DEVS world: CD++ (Wainer), COSMOS (Sarjoughian), Framworks in James (Urhmacher/Himmelspach), ...
  - OSA (Dalle): maven-based solution+ public repo
Thanks!
Benefits of Reproducibility

• A scientific "booster"
  - Use a previous experiment as a starting point new results
  - Possibly a long time later
    • Eg. maths theorem can be reproved years later
    • Eg. maths theorems can be reused for proving new theorems
• A first step toward better science practices
  - eg. Traceability: identify impacted/dependent publications
  - Invalidate/rerun scientific workflows
    • eg. MyExperiment/Taverna
Usage Examples & Applications

• Example 1: Network Performance Evaluation
  – Scientific study: assess/compare performance
  – Reproduce protocol/network operation
  – Can be Simulation, Emulation, or Hybrid

• Useful Levels of Reproducibility:
  – L1 : YES if parallel simulation
  – L2 : Yes, the least if Emulation/Hybrid
  – L3 : Yes (eg. for V&V, using another simulator/model implem.)
  – L4 : Marginal (eg. years later, based on published description, to check that results are still reproducible/valid)
Usage Examples & Applications

• Example 2: Road Traffic
  - Traffic management: identify/minimize congestion due to some events
  - OD matrix defines demands on network

• Useful Levels of Reproducibility:
  - L1: possibly, if parallel simulation
  - L2, L3, L4: no
Usage Examples & Applications

• Example 3: Training
  - Put users in situations
  - Predefined scenarios: eg. nuclear plant incident, ...

• Useful Levels of Reproducibility:
  - L1: yes if parallel simulation
  - L2: yes, learning based on repetition
  - L3: marginal use (training with multiple different simulators?)
  - N/A (interactivity vs. output)
Usage Examples & Applications

• Example 4: Gaming
  - Put users in situations
  - Predefined scenarios (levels)

• Useful Levels of Reproducibility:
  - L1: no (parallel simulation unlikely)
  - L2: yes, part of the game design
  - L3: Irrelevant (except maybe with board/card games?)
  - N/A (interactivity vs. output)
What is Reproducibility?

• In "old" sciences: defined by standards
  – eg. NIST in US about physics
  – specification with respect to results of measurement and experimental conditions

• In Computer Science: a research topic

• In Simulation?
A Possible Classification of Simulation Reproducibility Levels (Dalle, WSC 12)

4 Levels

- **L1**: Deterministic, identical computation
  - a.k.a. repeatability (in simulation)
  - Use the same program again
  - Produce the exact same data

- **L2**: Non-deterministic, identical computation
  - Relaxed form, eg. due to parallel execution or RT-coupling
  - Use the same program again
  - Most certainly lead to different data
A Possible Classification of Simulation Reproducibility Levels (Dalle, WSC 12)

- **L3** Identical scenario and instrumentation
  - Based on a detailed specification
  - Write a new program
  - May lead to different data

- **L4** Similar Scenario and Instrumentation
  - Based on a loose specification (similar vs. identical)
    - eg. 1000 nodes with random connections each running on-off CBR traffic source...
  - Build a scenario with same expected behavior as original experiment

**RR:** Repeatability (strong?)