Reproducibility in Computationally-enabled Research

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

The Judith Resnik Year of Women in ECE Seminar Series
Carnegie Mellon University
Pittsburgh, PA

April 6, 2016
Agenda

1. A Short History of Reproducibility
2. Defining Reproducibility
3. An Experimental Framework for Machine Learning
4. The “Whole Tale” Project
5. (if time) The Reproducible Research Standard
A Very Short History of Computational Reproducibility

“In 1990 I set for SEP [Stanford Exploration Project] a goal of reproducible documents and results.”

Jon Claerbout introduces the term “reproducible research” for computational results.
“Really Reproducible Research” (1992) inspired by Jon Claerbout:

“The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.” David Donoho, 1998

Note the difference between: reproducing the computational steps and replicating the experiments independently (including data collection, software implementation)
In Psychology

• 2011: Joseph Simmons, Leif Nelson, and Uri Simonsohn publish “False-positive psychology,” in Psychological Science and introduce the term “researcher degrees of freedom” and later “p-hacking”

• 2012: Daniel Kahneman’s letter exhorting psychology researchers to proactively respond to doubts regarding findings in priming studies appears in Nature.

• 2015: Center for Open Science publishes a replication project in Science for three psychology journals with success for .33 to .5 of the 100 articles attempted.
In Pre-Clinical Research

• 2011: Two publications by Amgen and Bayer claimed they could not reproduce the results from the majority of published studies they investigated (25% and 11% respectively).
The resulting fallout..

1. Computation as a tool for science.
   - *General thesis*: The use of computation to derive scientific findings has implications on reporting standards and research conduct to ensure verifiability and reproducibility of claims.
   - *Solution*: code, data, workflow release; full reporting of computational steps; care to design experimental computational environments that support reproducibility, and others.

   - *General thesis*: Statistical methods used for inference in the social sciences don't support the strength of claims being made.
   - *Solution*: Ensure better methods through preregistration, better training in statistics, use of power calculations prior to experimentation, and (to some degree) provide computational environments that support good statistical practice, and others.

3. Pre-clinical and health research irreproducibility.
   - *General description* (the cause is not agreed upon, in part because of how the problem came to light): demonstrated irreproducibility in published findings is a problem.
   - *Solution*: improve researcher incentives to conform to known and understood scientific standards and best practices.
Defining Reproducibility

“Computational Reproducibility”

“Statistical Reproducibility”

“Empirical Reproducibility”

V. Stodden, IMS Bulletin (2013)
Statistical Reproducibility

- False discovery, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.
- Low power, poor experimental design, nonrandom sampling,
- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,
- inappropriate tests or models, model misspecification,
- Model robustness to parameter changes and data perturbations,
- Investigator bias toward previous findings; conflicts of interest.
- …
Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hines,1,5* Ying Su,2,4,5• Irene Kuhn,1 Kornelia Polyak,2,4,5 and Mina J. Bissell1,5
1Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R255A, 1 Cyclotron Road, Berkeley, CA 94720, USA
2Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
3Department of Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA
4Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
5These authors contributed equally to this work
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.cellrep.2014.02.021

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other’s fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of an astonishing 50 individual markers per cell (Cheung and Ulz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.’s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.’s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

Reproducibility Issues in Research with Animals and Animal Models

The missing "F": Reproducibility in a Changing Research Landscape

A workshop of the Roundtable on Science and Welfare in Laboratory Animal Use

National Academy of Sciences, NAS 125
2100 C Street NW, Washington DC
June 4–5, 2014

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.
The Scientific Method

Traditionally two branches of the scientific method:

• Branch 1 (deductive): mathematics, formal logic,
• Branch 2 (empirical): statistical analysis of controlled experiments.

Many claim the emergence of new branches:

• Branch 3,4? (computational): large scale simulations / data driven computational science.
“It is common now to consider computation as a third branch of science, besides theory and experiment.”

“This book is about a new, fourth paradigm for science based on data-intensive computing.”
The Ubiquity of Error

The central motivation for the scientific method is to root out error:

• Deductive branch: the well-defined concept of the proof,

• Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation and Data Science present only potential third/fourth branches of the scientific method (Donoho et al. 2009), until the development of comparable standards.
Querying the Scholarly Record

• Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;

• Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;

• List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;

• Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;

• Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the year 2003 and list the trial name and histogram side by side.

Courtesy of Donoho and Gavish 2012
A “Experiment Definition” for Machine Learning Research


Step 1: Find articles in the literature that analyze the dataset (we found 30).

Step 2: Identify (and replicate) both feature selection / preparation steps and classification algorithms.
## Framework

<table>
<thead>
<tr>
<th>Article</th>
<th>Feature Selection Method</th>
<th>Classification Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ben-Dor et al (2000)</td>
<td>Threshold Number of Misclassifications (TNoM)</td>
<td>SVN (linear and quadratic), AdaBoost</td>
</tr>
<tr>
<td>Dudoit et al (2002)</td>
<td>Between Group / Within Group Variance</td>
<td>LDA, QDA, Classification Trees, Neural Nets</td>
</tr>
</tbody>
</table>
Work in progress..

Our findings so far:

• meaningful comparisons hindered by lack of consistent reference dataset,

• research on data preparation and feature selection limited by lack of transparency,

• differences in research pipeline impossible to discern without code published for reproducibility

• generalizability to other settings?

Need for a “Common Task Framework"
The “Common Task Framework"

Originated in the 1980’s in Natural Language Processing

1. A detailed “evaluation plan”
   • developed in consultation with researchers
   • and published as the first step in the project.

2. Automatic evaluation software
   • written and maintained by NIST
   • and published at the start of the project.

3. Shared data
   • Training and “dev(elopment) test” data is published at start of project;
   • “eval(uation) test” data is withheld for periodic public evaluations

Slide courtesy of Liberman 2015
Algorithm Transparency

In 2008 Google Flu Trends claimed it can tell you whether "the number of influenza cases is increasing in areas around the U.S., earlier than many existing methods."

In 2013 Google Flu Trends was predicting more than double the proportion of doctor visits for flu than the CDC.

Today:

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of flu and dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we’re excited to see what comes next. Academic research groups interested in working with us should fill out this form.

Sincerely,
The Google Flu and Dengue Trends Team.
What Happened?

- How did Google Flu Trends work? What was the data collection process? What was the algorithm?

- Why should we believe Google Flu Trends output? Many people did in 2008..

---

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

In February 2013, Google Flu Trends (GFT) made headlines. The algorithm in 2009, and this model has run ever since, with a
Encouraging Reproducibility While Expanding Access to Massive Computation

We see the convergence of two (ordinarily antagonistic) trends.

- Scientific projects will become massively more computing intensive

- Scientific computing dramatically more transparent
“Experiment Definition Systems”

- Define and create “Experiment Definition Systems” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis.

- The two trends need to be addressed simultaneously: better transparency will allow people to run much more ambitious computational experiments. And better computational experiment infrastructure will allow researchers to be more transparent.
Infrastructure Solutions

Research Environments

- Verifiable Computational Research
  - knitR
- Collage Authoring Environment
  - Sumatra
  - Galaxy
- SHARE
  - Sweave
- SOLE
  - GenePattern
  - torch.ch
- Code Ocean
  - Cyverse
- Open Science Framework
  - IPOL
  - Whole Tale
- Jupyter
  - NanoHUB
- Vistrails
  - Popper

Workflow Systems

- Taverna
- Wings
- Pegasus
- CDE
  - Reprozip
- Collage Authoring Environment
  - Sumatra
  - Galaxy
- Open Science Framework
  - IPOL
  - Whole Tale

Dissemination Platforms

- ResearchCompendia.org
- DataCenterHub
- RunMyCode.org
- ChameleonCloud
- Occam
- RCloud
- TheDataHub.org
- Madagascar
Whole Tale: Merging Science and Cyberinfrastructure Pathways

Bertram Ludaescher, Kyle Chard, Niall Gaffney, Matthew B. Jones, Jaroslaw Nabrzyski, Victoria Stodden, Matt Turk

wholetale.org
Whole Tale: What’s in a name?

- (1) Whole Tale ⇔ Whole **Story**:
  - **Support** (computational & data) **scientists**
  - … along the **complete research lifecycle**
  - ... from **experiment** to (new kind of) **publication**
  - ... and back!

![Diagram showing the research lifecycle and the role of cyberinfrastructure.](Image)
Whole Tale: What’s in a name?

- (2) Whole Tale ↔ For the Long Tail of Science
  - “Big data & compute for mere mortals”

Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogeneous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich variety of specialty research data sets. The extreme right portion of the long tail includes data that are unpublished; such as siloed databases, null findings, laboratory notes, animal care records, etc. These dark data hold a potential wealth of knowledge but are often inaccessible to the outside world.
Whole Tale Vision

- The Old Way:
  - Scholarly Publication \(\leftrightarrow\) Data \(\leftrightarrow\) Code

- The Emerging Way:
  - Scholarly Publication \(\leftrightarrow\) Data \(\leftrightarrow\) Code

- The New Way:
  - “Living” Publication \(\leftrightarrow\) Data \(\leftrightarrow\) Code
    \(=\) Computational Narrative
  - (more easily) Reproducible Science

.. participate in and share the experience of inquiry
Problems Facing Researchers

Workflow for data research is **fragmented**:

- Data comes from many sources and is "**integrated the old fashioned way**" *(email, Excel, …)*

- Use cloud services **copying data** from *(Drop)*Box, Google-Drive, … to local storage with a distributed directory structures to organize (and provide discovery) to data

- Data provenance is **not captured** *(custom scripts, some version of a community developed and supported codebase)*

- Publication of data with link to publication *(never mind DOIs, DMP)* is **not sufficient for reproducibility**
So what do we do about this?

- WT will leverage & contribute to **existing CI and tools** to support the **whole science story** (= run-to-pub-cycle), and providing access to big data via CI and compute for **long tail** researchers.

  ➡ **Integrate tools to simplify usage and promote best practices**

- **NSF CC*DNI DIBBS:**
  - 5 Institutions, 5 Years ($5M total)
  - Cooperative Agreement
Specific Goals of Whole Tale

- Expose existing CI
  - … through popular frontends (Jupyter, RStudio, ..)

- Develop necessary “software glue”
  - … for seamless access to different CI-backend capabilities

- Enhance data-to publication lifecycle
  - … by empowering scientists to create computational narratives in their usual programming environments
Iterative Design through Working Groups

Merging Science & CI Pathways … through Working Groups

Working Groups (Science Drivers)
- Astronomy and Astrophysics
- Earth & Env. Sciences, Archaeology
- Bioinformatics & Genomics
- Materials Sciences
- Social Sciences

Working Groups (CI Providers)
- Tools Development
- Reproducibility
- Information Science
- Education and Training

Working Groups Driving Use Cases and Adoption

Iterative Design

Working Groups to Provide Key Components
Computational Barriers to Reproducibility

• rerunning same code, same parameter settings, same system can produce different results (?),

• same code, but updated libraries, compiler, operating system, etc,

• software customization to underlying architectures; portability, modularity, re-usability,

• numerical stability of the underlying software,

• unique hardware, scarce allocations, long runtimes..
Infrastructure Stability

- Joint work with my post-doc Matthew Krafczyk

- Sources of variability from the software infrastructure:
  - Enzo galaxy simulation code on Blue Waters
  - Sensitivity of scientific results to infrastructure/environment parameter changes (compiler, libraries, threading, etc.)
This workshop will consider ways to make code and modeling information more readily available, and include a variety of stakeholders.

The computational steps that produce scientific findings are increasingly considered a crucial part of the scholarly record, permitting transparency, reproducibility, and re-use. Important information about data preparation and model implementation, such as parameter settings or the treatment of outliers and missing values, is often expressed only in code. Such decisions can have substantial impacts on research outcomes, yet such details are rarely available with scientific findings.

http://www.aaas.org/event/iii-arnold-workshop-modeling-and-code
Feb 16-17, 2016
REPRODUCIBILITY

Enhancing reproducibility for computational methods
Data, code, and workflows should be available and cited

By Victoria Stodden,1 Marcia McNutt,2 David H. Bailey,3 Ewa Deelman,4 Yolanda Gil,4 Brooks Hanson,5 Michael A. Heroux,6 John P.A. Ioannidis,7 Michela Taufer8

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness Promotion (TOP) guidelines (1) and recommendations for field data (2), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers representing different domains. The goal is to understand how computational results were derived and to reconciling any differences that might arise between independent replications (4). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (5, 6), which includes workflow information that explains what raw data and intermediate results are input to which computations (7). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS
Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or scripts.

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fWjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. All data, code, and workflows, including software written by the authors, should be cited in the references section (10). We suggest that software citation include software version information and its unique identifier in addition to providing data citation information.
Workshop Recommendations: “Reproducibility Enhancement Principles”

RECOMMENDATION 1: To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.

RECOMMENDATION 2: To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

RECOMMENDATION 3: To enable credit for shared digital scholarly objects, citation should be standard practice.

RECOMMENDATION 4: To facilitate reuse, adequately document digital scholarly artifacts.
Workshop Recommendations: “Reproducibility Enhancement Principles”

RECOMMENDATION 5: Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.

RECOMMENDATION 6: Use Open Licensing when publishing digital scholarly objects.

RECOMMENDATION 7: To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.
Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) via the Constitution and subsequent Acts:

“To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.” (U.S. Const. art. I, §8, cl. 8)

**Argument**: both types of intellectual property are an imperfect fit with scholarly norms, and require action from the research community to enable re-use, verification, reproducibility, and support the acceleration of scientific discovery.
Copyright

- Original expression of ideas falls under copyright by default (papers, code, figures, tables..)

- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original

- limited time: generally life of the author +70 years

- Exceptions and Limitations: e.g. Fair Use.
Patents

Patentable subject matter: “new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof” (35 U.S.C. §101) that is

1. **Novel**, in at least one aspect,

2. **Non-obvious**,

3. **Useful**.

USPTO Final Computer Related Examination Guidelines (1996) “A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena” (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).
Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.
Legal Issues in Data

• In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist PublIns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).

• Copyright adheres to raw facts in Europe.

• the possibility of a residual copyright in data (attribution licensing or public domain certification).

• Legal mismatch: What constitutes a “raw” fact anyway?
Privacy and Data

• HIPAA, FERPA, IRB mandates create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/ )

• Potential privacy implications for industry generated data.

• Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..
Ownership: What Defines Contribution?

• Issue for producers: credit and citation.

• What is the role of peer-review?

• Repositories adding meta-data and discoverability make a contribution.

• Data repositories may be inadequate: velocity of contributions

• Future coders may contribute in part to new software, other software components may already be in the scholarly record. Attribution vs sharealike.

  ➡ (at least) 2 aspects: legal ownership vs scholarly credit.

• Redefining plagiarism for software contributions.
Licensing in Research

Background: Open Source Software

Innovation: Open Licensing

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

Hundreds of open source software licenses:

- GNU Public License (GPL)
- (Modified) BSD License
- MIT License
- Apache 2.0 License
- ... see http://www.opensource.org/licenses/alphabetical
The Reproducible Research Standard

The *Reproducible Research Standard (RRS)* (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

- Remove copyright’s barrier to reproducible research and,
- Realign the IP framework with longstanding scientific norms.
Conclusions

We see the convergence of two (ordinarily antagonistic) trends:

➡ Scientific projects will become massively more computing intensive

➡ Research computing will become dramatically more transparent

Reinforcing trends, resolution essential for verifying and comparing findings.
Approach

WT will integrate established CI components, creating a simple, unified environment to use, share, and publish data and workflows:

1. **Unified Authentication** via Globus Auth
2. **Abstracted Storage** Layer with a unified namespace
3. **Integrated** Python and R APIs with **Jupyter Notebook Environments**
4. **Ingest and publication** service linking data, computations, and scholarly articles
5. **NextCloud integration** for “Dropbox like interface”
6. **Event System** to react to changes (e.g. new data published)
7. **Data Dashboard** to ease data management and analysis

→ Capture full workflow via notebooks, scripts, and applications to be published along with data and research publications
Whole Tale Identity, Authentication and Authorization Landscape

Identities

Cyberinfrastructure
Result and Artifact Review and Badging

An experimental result is not fully established unless it can be independently reproduced. A variety of recent studies, primarily in the biomedical field, have revealed that an uncomfortably large number of research results found in the literature fail this test, because of sloppy experimental methods, flawed or unknown data analysis, or both. Publication of results is a necessary component of advancing the science, but their value is enhanced by the ability of other researchers to reproduce the results.
Terminology.

A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology (VIM); see the Appendix for details.

- **Repeatability** (Same team, same experimental setup)
  
  The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

- **Replicability** (Different team, same experimental setup)
  
  The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author’s own artifacts.

- **Reproducibility** (Different team, different experimental setup)
  
  The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.
Proposition 1

• We propose a major effort to develop a new infrastructure that promotes good scientific practice downstream like transparency and reproducibility.

• But plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work. Enables efficiency and productivity, and discovery.
Inducing a Reproducibility Industry by Grant Set-asides

- Previously, NIH required that clinical trials hire Biostatistician PhD's to design and analyze experiments. This set-aside requirement more or less directly transformed clinical trials practice and resulted in much more good science being done. It also spawned the modern field of Biostatistics, by creating a demand for a specific set of services and trained people who could conduct them.

- Why not try a similar idea for reproducibility?
Proposition 2

• We propose a set-aside that will initiate and sustain near-universal following of reproducibility and transparency practices.

• The set-aside has these components:

  1. Each grant contains a budget of X times number of anticipated papers. (say X=$500)

  2. These funds are only to be used to pay for reproducibility certification by an accredited reproducibility certifier.

  3. The certifier *both* provides a "badge" on the publication, *and* hosts digital content allowing partial or complete reproduction of results of the paper.
4. A certifier might be a traditional entity, like a journal, a society, or a library. The certifier might also be a novel institution.

5. The certifier would develop and disseminate software tools that potential certifiees would use during their research. These entities will also determine standards for research areas.

6. It is to be expected although not a requirement, that reproducibility certification would become automated, although in early days it might be humans certifying reproducibility.

Contrast with Reproducibility Editors being used by some journals.
3. Looking Ahead

Ways forward:

1. Encouraging Reproducibility While Expanding Access to Massive Computation

2. Inducing a “Reproducibility Industry” by Grant Set-asides

*Joint work with David Donoho*
Encouraging Reproducibility While Expanding Access to Massive Computation

We see the convergence of two (ordinarily antagonistic) trends.

- Scientific projects will become massively more computing intensive

- Scientific computing dramatically more transparent
“Experiment Definition Systems”

- Define and create “Experiment Definition Systems” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis.

- The two trends need to be addressed simultaneously: better transparency will allow people to run much more ambitious computational experiments. And better computational experiment infrastructure will allow researchers to be more transparent.
Proposition 1

• We propose a major effort to develop a new infrastructure that promotes good scientific practice downstream like transparency and reproducibility.

• But plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work. Enables efficiency and productivity, and discovery.
Inducing a Reproducibility Industry by Grant Set-asides

• Previously, NIH required that clinical trials hire Biostatistician PhD's to design and analyze experiments. This set-aside requirement more or less directly transformed clinical trials practice and resulted in much more good science being done. It also spawned the modern field of Biostatistics, by creating a demand for a specific set of services and trained people who could conduct them.

• Why not try a similar idea for reproducibility?
Proposition 2

• We propose a set-aside that will initiate and sustain near-universal following of reproducibility and transparency practices.

• The set-aside has these components:

1. Each grant contains a budget of X times number of anticipated papers. (say X=$500)

2. these funds are only to be used to pay for reproducibility certification by an accredited reproducibility certifier.

3. the certifier *both* provides a "badge" on the publication, *and* hosts digital content allowing partial or complete reproduction of results of the paper.
4. A certifier might be a traditional entity, like a journal, a society, or a library. The certifier might also be a novel institution.

5. The certifier would develop and disseminate software tools that potential certifiees would use during their research. These entities will also determine standards for research areas.

6. It is to be expected although not a requirement, that reproducibility certification would become automated, although in early days it might be humans certifying reproducibility.

Contrast with Reproducibility Editors being used by some journals.
Conclusion

- Reproducibility issues travel will all computational research.
- Computation is near-ubiquitous in modern research.
- Cyberinfrastructure is underdeveloped and could help resolve irreproducibility.
- Grant set-asides could foment a “Reproducibility Industry”
A Credibility Crisis

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren’t even true.

By Michael Hilzik
October 27, 2013

In today’s world, brimful as it is with opinion and falsehoods masquerading as facts, you’d think the one place you can depend on for verifiable facts is science.

You’d be wrong. Many billions of dollars’ worth of wrong.