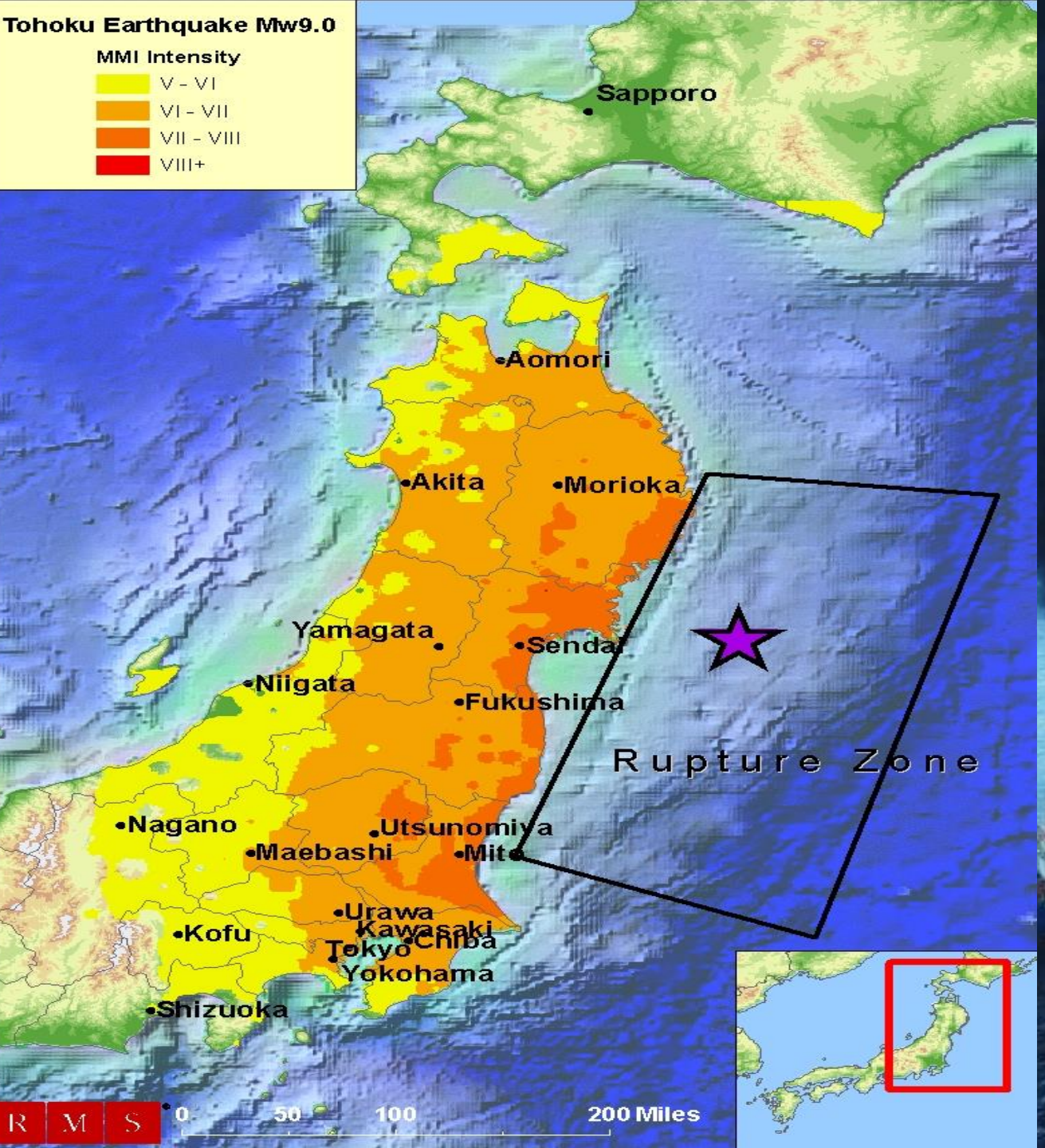
Data sharing: some cultural perspectives

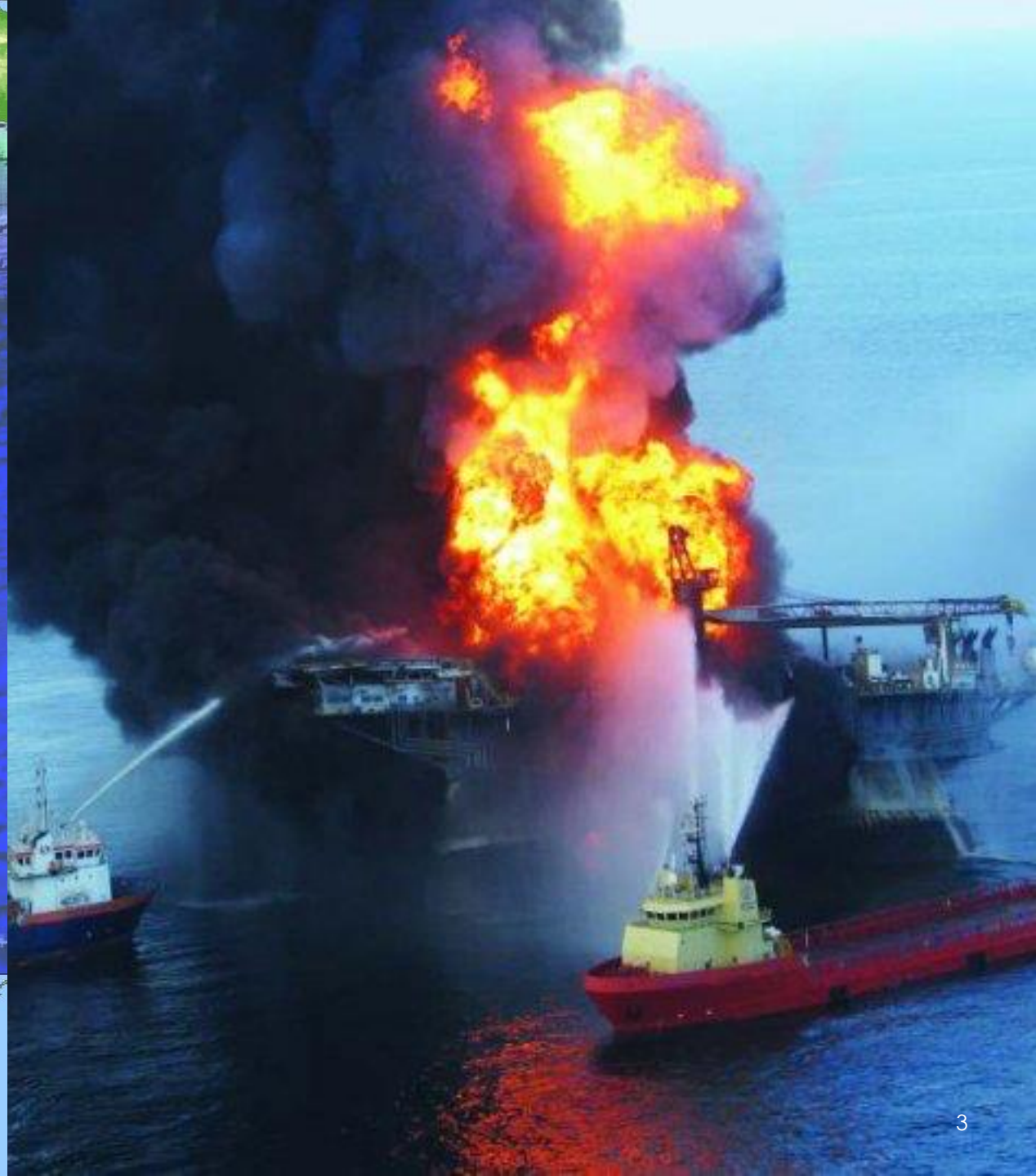
MARCIA MCNUTT
PRESIDENT

NATIONAL ACADEMY OF SCIENCES

Data sharing is centerpiece of open science

- Enables reuse of data for new purposes
- Promotes trust in science from non-scientists
- Facilitates verification of findings by reviewers and readers
  - But not all reproducible research is correct;
  - And not all irreproducible research is wrong;
  - Extension of results and verifying findings through completely independent methods are time-honored approaches


Tohoku Earthquake Mw9.0
MMI Intensity
V - VI
VI - VII
VII - VIII
VIII+
Sapporo
Aomori
Akita
Morioka
Yamagata
Sendai
Niigata
Fukushima
Rupture Zone
Nagano
Utsunomiya
Maebashi
Mito
Urawa
Kawasaki
Kofu
Chiba
Tokyo
Yokohama
Shizuoka
0
50
100
200 Miles
R
M
S

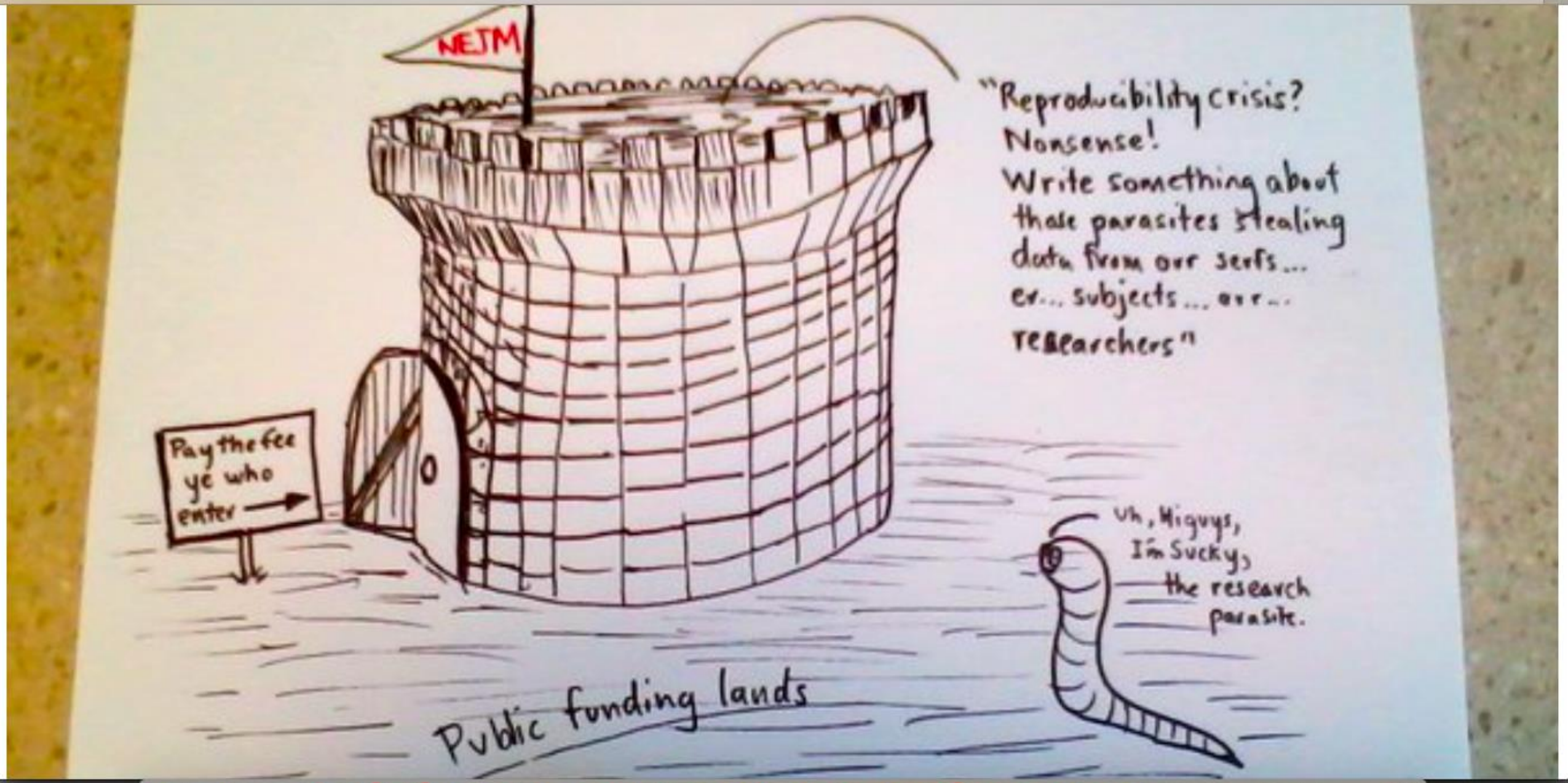


Data sharing cultures vary with field

- Some disciplines have extensive histories of sharing their data, e.g.,
  - Astronomy
  - Oceanography
  - High-energy physics
- Other disciplines are working to counteract proprietary attitudes towards information

Medical Journal on Data Sharing

- Encouraging sharing will create "data parasites"
- "Someone else might be using my data only to disprove my results."
- Those not involved in the original research will not know how to use the data
- **Better approach: Collaborate with original authors – invite them to be coauthors on the study as a condition on using their data.**


NEJM
Pay the fee ye who enter
"Reproducibility crisis? Nonsense! Write something about those parasites stealing data from our serfs... er... subjects... err... researchers"
Uh, Higuys, I'm Sucky, the research parasite.
Public funding lands

SHARE

160

6

EDITORIAL

#IAmAResearchParasite

Marcia McNutt

+ Author Affiliations

Science 04 Mar 2016:
Vol. 351, Issue 6277, pp. 1005
DOI: 10.1126/science.aaf4701

PHOTO: STACEY PENTLAND PHOTOGRAPHY

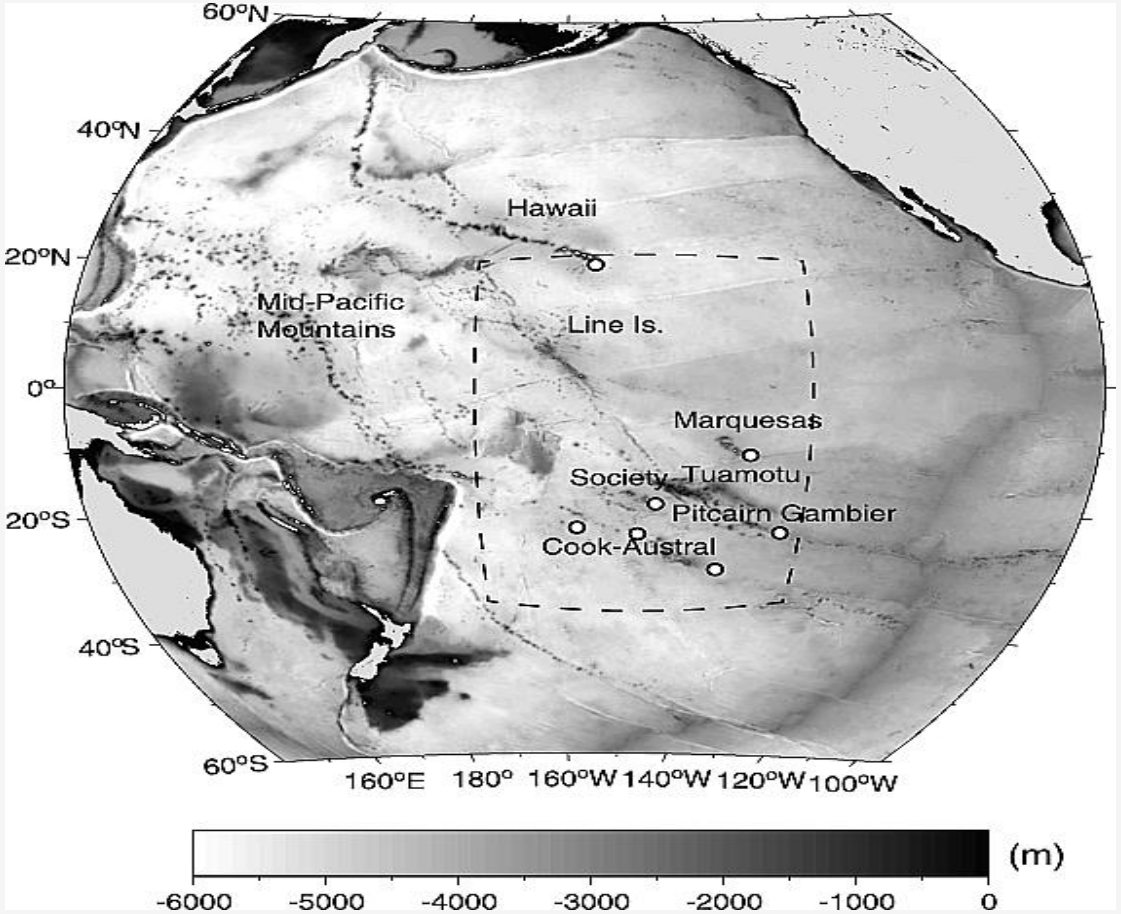
In the midst of steady progress in policies for data sharing, a recent editorial expressed a contrarian view.* The authors described the concern of some scientists about the rise of an underclass of "research parasites" who exploit data sets that are collected and curated by others. Even worse, these parasites might use such data to try to disprove the conclusions posited in the data's original source studies. The editorial raised the points of how anyone not involved in the original study could use the data without misrepresenting it, and the danger of perhaps arriving at erroneous conclusions. The editorial advised instead that data sharing be implemented by involving the authors of the original study as coauthors in follow-up research. The research community immediately took to Twitter under the hashtag #IAmAResearchParasite to voice opposition to the editorial.

> ***"There are costs...for re-collecting data for new uses."***

PHOTO: ©HARDIE/ALAMY

Much of what we know about the large-scale features of this planet is apparent thanks to widespread data-sharing practices and the early establishment of data banks in the geosciences. Aspects such as determining the shape of the ocean floor, ocean chemistry, the internal structure of Earth's deep interior, the physics and chemistry of the atmosphere, and many other topics could not have been ascertained from a single investigator's field program. One meta-analysis I published on the South Pacific benefited from observations of my own and those of others, including the 18th-

Hard to Include Everyone as Co-Authors



Journal of Geophysical Research: Solid Earth
Volume 110, Issue B9, B09408, 20 SEP 2005 DOI: 10.1029/2004JB003465
http://onlinelibrary.wiley.com/doi/10.1029/2004JB003465/full#jgrb14365-fig-0001

Importance of Reproducing Results

SHARE

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

+ Author Affiliations

↵†Corresponding author. E-mail: nosek@virginia.edu

Science 28 Aug 2015:
Vol. 349, Issue 6251, pp.
DOI: 10.1126/science.aac4716

Article | Figures & Data | Info & Metrics | eLetters | PDF

Empirically analyzing empirical evidence

One of the central goals in any scientific endeavor is to understand causality. Experiments that seek to demonstrate a cause/effect relation most often manipulate the postulated causal factor. Aarts *et al.* describe the replication of 100 experiments reported in papers published in 2008 in three high-ranking psychology journals. Assessing whether the replication and the original experiment yielded the same result according to several criteria, they find that about one-third to one-half of the original findings were also observed in the replication study.

Science, this issue 10.1126/science.aac4716

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

What Journals are Doing

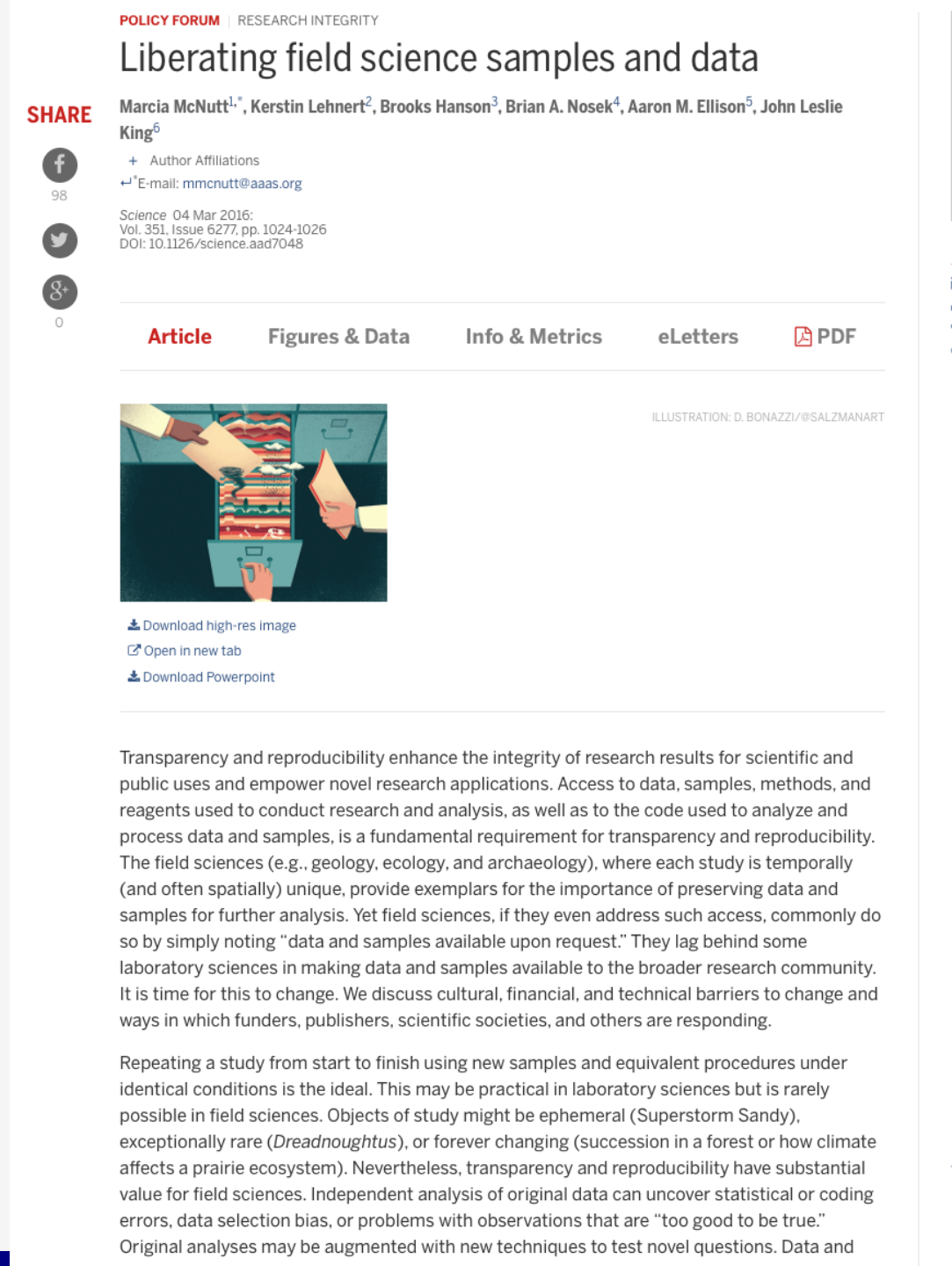
- TOP (Transparency and Openness Promotion) standards published in Science
- Now more than 500 journals have signed on
- Include openness standards for samples and data

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

Motivations vary by stakeholder…

- Funders want to see value -> require deposit in accessible repositories & measure reuse
- Scientists need to get credit: data hoarding -> citations (with persistent identifiers, e.g., DOI's)
- Repositories want to streamline data flow to support communities -> promote development of aps for deposit/withdrawal
- Journals want to promote quality/reproducibility -> provide automated alert for deposit
- Support community efforts to create metadata
- Develop careers for data professionals

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine



POLICY FORUM | RESEARCH INTEGRITY

Liberating field science samples and data

SHARE 98 0

Marcia McNutt[1,*], Kerstin Lehnert[2], Brooks Hanson[3], Brian A. Nosek[4], Aaron M. Ellison[5], John Leslie King[6]

+ Author Affiliations

↵*E-mail: mmcnutt@aaas.org

Science 04 Mar 2016:
Vol. 351, Issue 6277, pp. 1024-1026
DOI: 10.1126/science.aad7048

Article | Figures & Data | Info & Metrics | eLetters | PDF

ILLUSTRATION: D. BONAZZI/@SALZMANART

Download high-res image
Open in new tab
Download Powerpoint

Transparency and reproducibility enhance the integrity of research results for scientific and public uses and empower novel research applications. Access to data, samples, methods, and reagents used to conduct research and analysis, as well as to the code used to analyze and process data and samples, is a fundamental requirement for transparency and reproducibility. The field sciences (e.g., geology, ecology, and archaeology), where each study is temporally (and often spatially) unique, provide exemplars for the importance of preserving data and samples for further analysis. Yet field sciences, if they even address such access, commonly do so by simply noting "data and samples available upon request." They lag behind some laboratory sciences in making data and samples available to the broader research community. It is time for this to change. We discuss cultural, financial, and technical barriers to change and ways in which funders, publishers, scientific societies, and others are responding.

Repeating a study from start to finish using new samples and equivalent procedures under identical conditions is the ideal. This may be practical in laboratory sciences but is rarely possible in field sciences. Objects of study might be ephemeral (Superstorm Sandy), exceptionally rare (*Dreadnoughtus*), or forever changing (succession in a forest or how climate affects a prairie ecosystem). Nevertheless, transparency and reproducibility have substantial value for field sciences. Independent analysis of original data can uncover statistical or coding errors, data selection bias, or problems with observations that are "too good to be true." Original analyses may be augmented with new techniques to test novel questions. Data and

Data Sharing Is Not Cheap

- Fund data repositories
- Support data professionals
- Create apps that automate data deposition and extraction
- Educate all in importance of quality control

Data Sharing is Not Easy

- Establish metadata standards
- Culture changes to:
  - Relinquish ownership of data
  - Start treating data as citable objects

Why We Must Implement

- Irreproducible research costs society
- Recollecting data that colleagues are not sharing takes funds away from worthy projects
- Data from ephemeral or temporally variable processes cannot be recreated later if not properly archived
- Collaborating with original authors is not always possible or practical, slows down the progress of science, and can bias results