CIS 602-01: Computational Reproducibility

Data Availability

Dr. David Koop
What are data?

Marie Curie’s notebook

Pisa Griffin

http://www.census.gov/population/cen2000/map02.gif

D. Koop, CIS 602-01, Fall 2016

[http://onlineda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php]

[C. Borgman, 2015]
"Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship."

- C. Borgman
Categories of Data

• Observational: e.g. weather, may go across time and location
• Computational: data from computer model and simulation
• Experimental: lab or field experiment (may be replicated)
• Records: e.g. government records

• (via National Science Board)

[C. Borgman, 2012]
Science and Data

Engineering researcher: “Temperature is temperature.”

Biologist: “There are hundreds of ways to measure temperature. ‘The temperature is 98’ is low-value compared to, ‘the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.’ That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.”

[C. Borgman, 2015]
The method is slow and too insensitive to distinguish between human and animal sources of bacteria. The more sophisticated method is quantitative polymerase chain reaction (qPCR), adapted from medical applications, which requires greater expertise and is much more expensive. This method is faster and more sensitive, but results will vary between laboratories due to choices of local protocols, filter material, machine type and model, and handling methods. Protocols and results are shared between partner laboratories seeking to perfect the method, but little other than the methods of data collection, protocols, and final curves might be reported in the journal articles. Biological samples are fragile; they degrade quickly or are destroyed in the analysis process.

At the other end of the specificity dimension are observatories, which are institutions for the observation and interpretation of natural phenomena. Examples include NEON and LTER in ecology (National Ecological Observatory Network, 2010; U.S. Long Term Ecological Research Network, 2010; Porter, 2010), GEON in the earth sciences (GEON, 2011; Ribes & Bowker, 2008), and synoptic sky surveys in astronomy (Panoramic Survey Telescope & Rapid Response System, 2009; Large Synoptic Sky Telescope, 2010; Sloan Digital Sky Survey, 2010). Observatories attempt to provide a comprehensive view of some whole entity or system, such as the earth or sky. Global climate modeling, for example, depends upon consistent data collection of climate phenomena around the world at agreed upon times, locations, and variables (Edwards, 2010).

The value of observatories lies in systematically capturing the same set of observations over long periods of time. Astronomical observatories are massive investments, intended to serve a large community. Investigators and others can mine the data to ask their own questions or to identify bases for comparison with data from other sources. Studies of the role of dust emission in star formation make use of observatory data. In this star dust scenario, a team of astrophysics researchers queries several data collections that hold observations at different wavelengths, extracting many years of observations taken in a specific star-forming region of interest. They apply several new methods of data analysis to model physical processes in star formation. By combining data from multiple observatories, they produce empirical results that enable them to propose a new theory. Typically the combined dataset is released when they publish the journal article describing their results.

Scope of data collection.

The second dimension of Figure 1 is the scope of data collection. At one pole are exploratory observatories, which focus on descriptive phenomena, and at the other pole are theoretical observatories, which focus on model systems. In between are empirical observatories, which focus on the collection of empirical data. The goal of research is the third dimension, with observational research at one pole and theoretical research at the other.
Approaches in Handling Data

- **People Involved**
  - Collaborative Teams
  - Individual Investigator

- **Labor to Collect Data**
  - By Hand
  - By Machine

- **Labor to Process Data**
  - By Hand
  - By Machine

**Scenarios:**
- BQ: Beach Quality
- SD: Star Dust
- OS: Online Survey
- AR: Archival Records

[C. Borgman, 2012]
Rationales for Sharing Data

The arguments dimension (vertical axis) positions the rationales based on whether they are research-driven or public-driven. The benefits dimension (horizontal axis) positions the rationales based on whether they are data producer-driven or data user-driven.

**Rationales**
- R1: Reproduce/verify
- R2: Serve public interest
- R3: Ask new questions
- R4: Advance research

**Arguments for Sharing**

Data Producers
- R1
- R4

Data Users
- R2
- R3

[C. Borgman, 2012]
Reproducibility

• "…it is the most problematic rationale for sharing research data"!
• Data is not enough
• Cannot reduce research to "mechanistic procedures"
• [Depends more on interpretation than data]

[C. Borgman, 2012]
Data Sharing

- "Difficult"
- Authors need to see use or value
- Data may rely on software, often cannot just publish the data
- How do we deal with streaming data?

[C. Borgman, 2012]
Lack of incentives to share data

- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality

[C. Borgman, 2015]
Data Reuse Scenarios

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries

[C. Borgman, 2015]
Lack of incentives to reuse data

- Identify useful data
  - Documentation
  - Interpretation
  - Software
- Cleaning
- Trust
- Credit
- Licensing

[C. Borgman, 2015]
Mandated data archiving greatly improves access to research data

How is data sharing mandated?

- Governments
- Funding agencies
- Journals
Procedure

- Find data from papers that use genetic analysis program Structure
- Data often archived in Dryad repository
- Group results based on the publication's policy on data availability
- 229 papers
## Results

### Table 1. Number of eligible articles per journal and number for which data were available online

<table>
<thead>
<tr>
<th>Policy</th>
<th>Journal</th>
<th>Eligible articles</th>
<th>Data online</th>
</tr>
</thead>
<tbody>
<tr>
<td>No policy</td>
<td>Conservation Genetics</td>
<td>47</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Crop Science</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Genetica</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TAG</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Recommend data archiving</td>
<td>BMC Evolutionary Biology</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>BJLS</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Journal of Heredity</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PLoS One</td>
<td>51</td>
<td>6</td>
</tr>
<tr>
<td>Mandatory data archiving</td>
<td>Journal of Evolutionary Biology</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Evolution</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Heredity</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Molecular Ecology</td>
<td>28</td>
<td>27</td>
</tr>
</tbody>
</table>
Since this is a logistic model, we can readily calculate the effect that the different policy types have on the likelihood that the data will be available. We explore these odds for each type of policy below, using “no policy” as the baseline.

Having a “recommend archiving” policy made it 3.6 times more likely that the data were online compared to having no policy. However, the 95% CI overlapped with 1 (0.96–13.6); hence, this increase in the odds is not significant. Overall, recommending data archiving is only marginally more effective than having no policy at all.

The data were 17 times more likely to be available online for journals that had adopted a mandatory data archiving policy but did not require a data accessibility statement in the manuscript. This odds ratio was significantly (95% CI: 3.7–79.6).

For “mandate archiving” journals where a data accessibility statement is required in the manuscript, the odds of finding the data online were 974 times higher compared to having no policy. The 95% CI on these odds is very wide (97.9–9698.8), but nonetheless shows that the combination of a mandatory policy and an accessibility statement is much more effective than any other policy type.

REQUESTING DATA DIRECTLY FROM AUTHORS

A number of the “recommend archiving” policies state that the data should also be freely available from the authors by request (see the Journal Policies file at doi: 10.5061/dryad.6bs31); hence, we wanted to evaluate whether obtaining data directly from authors is an effective approach. Part of the dataset collection for our reproducibility study (5) involved e-mailing authors of papers from two of the “recommend archiving” journals (BMC Evolutionary Biology and PLoS One) and requesting their structural input files. Here, we examine how often these requests led to us obtaining the data. We did not e-mail the authors of articles where the data were already available online. A detailed description of our data request process appears on Dryad (doi: 10.5061/dryad.6bs31), but we essentially contacted corresponding and senior authors of each article up to 3 times over a 3-wk period, and recorded if and when the data were received.

We obtained data directly from the authors for 7 of the 12 eligible articles in BMC Evolutionary Biology, and 27 datasets from 45 articles from PLoS One (Table 1). All seven of the BMC Evolutionary Biology datasets arrived between 8 and 14 d after our initial request. Ten of the PLoS One datasets came within 1 wk, 13 came between 8 and 14 d, and 4 arrived between 15 and 21 d. Unlike the online data, which could generally be obtained within a few minutes, the requested datasets took a mean of 7.7 d to arrive, with one author responding that the dataset had been lost in the year since publication. More than one e-mail had to be sent to the corresponding and/or senior author for 53% of papers, and the authors of 29% of the papers did not respond to any of our requests. No data were received 21 d after our initial request. We also note that requesting data via e-mail did upset some authors, particularly when they were reminded of the journal’s data archiving policy or when multiple e-mails were sent.

Our average return of 59% in an average of 7.7 d is markedly better than has been reported in similar studies: Wicherts et al. (8) received only 26% of requested datasets after 6 mo of effort with authors of 141 psychology articles, and Savage and Vickers (9) received only 1 of 10 eligible papers with data available online.
Differences in Policy

• PLOS One has a very long-winded, more nuanced policy:
  - "data should be provided in an open access institutional repository, a general data repository such as Dryad, or as Supporting Information files with the published paper"
• BMC Evolutionary Biology: data should be freely available
• Why? Domain differences
Nature data availability and data citations

• New policy: July 2016
• http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf
The availability of research data declines rapidly with article age

Study data availability

- Request data from specific types of articles
- 516 articles
- 1991 to 2011

Find working email and send message requesting data
- May have multiple addresses
- May not work
- May not respond

Use logistic regression to evaluate various criteria
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data.

Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it.

Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those published thereafter.

### Table 1. Breakdown of Data Availability by Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
<td>36</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>8 (23%)</td>
<td>35</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>7 (23%)</td>
<td>30</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>6 (15%)</td>
<td>40</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>8 (19%)</td>
<td>43</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>11 (24%)</td>
<td>46</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>13 (29%)</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>17 (26%)</td>
<td>66</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>18 (26%)</td>
<td>69</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>30 (38%)</td>
<td>80</td>
</tr>
<tr>
<td>Totals</td>
<td>131 (25%)</td>
<td>194 (38%)</td>
<td>33 (6%)</td>
<td>37 (7%)</td>
<td>20 (4%)</td>
<td>101 (19%)</td>
<td>121 (23%)</td>
<td>516</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages. The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those published afterward. The data availability for the most recent years was higher than that for the earlier years, which can be seen in Table 1. The breakdown of data availability by year of publication is as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working Email</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>36</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>35</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>30</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>40</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>43</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>46</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>66</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>69</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>80</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.
Number of Emails in Paper

The figure shows the number of emails in a paper as a function of the age of the paper in years. The trend is downward, indicating a decrease in the number of emails over time. The red dots represent individual data points, while the black line is a fitted regression curve. The shaded area around the line indicates the confidence interval for the regression model.

The x-axis represents the age of the paper in years, ranging from 5 to 20. The y-axis represents the number of emails in the paper, ranging from 0.0 to 1.5.

The title of the figure is "Number of Emails in Paper."
Data Availability Declines with Article Age

Figure 2. The Effect of Article Age on the Number of E-mails in Articles Published After 2000

- The proportion of e-mails from the paper that appeared to work against the data
- The number of e-mails in articles published after 2000
- The larger number of e-mails in recent papers may mean that the issue of e-mails in articles published after 2000
- Researchers in e.g., 2031 will be able to try a wider range of e-mails.

Indicators of the Effect of Article Age on the Number of E-mails

- Figure 2A: The proportion of e-mails that worked against the data
- Figure 2B: The number of e-mails that worked against the data
- Figure 2C: The number of e-mails found by searching on the web
- Figure 2D: Predicted probability that an individual e-mail worked against the data

The diagram shows a trend where the percentage of e-mails that worked decreases with the age of the paper. The data is represented with red dots and a trend line indicating the decrease in the proportion of working e-mails as the age of the paper increases.
Number of Emails Found on Web

![Graph showing the number of emails found on the web against the age of the paper. The graph includes a trend line and shaded area representing the 95% confidence interval. The x-axis represents the age of the paper (in years), ranging from 5 to 20, and the y-axis represents the number of emails found on the web, ranging from 0.0 to 1.2. The data points are indicated by red dots, and the black line represents the predicted probability from a logistic regression model. The shaded area around the line indicates the 95% confidence interval for the predicted values.]
Web Email Worked

Figure 2. The Effect of Article Age on the Number of E-mails in Articles Published After 2000

The proportion of e-mails from the paper that appeared to work on the web against article age.

The line indicates the predicted probability from a Poisson (A and C) or logistic (B and D) regression, and the gray area shows the 95% CI of this estimate.

The graph shows a trend where the percentage of web emails that worked decreases as the age of the paper increases, indicating that older papers have a lower chance of email addresses being active.

Overall, we only received 19.5% of the requested data sets, and only 11% for articles published before 2000. We must have used in a DFA. Since this study is focused on the data that still existed dropped from 100% in 2011 to 33% in 1991. We therefore also set the condition that the data that has been found for these data or when they would become impossible to know what uses would have been found for these data or when they would become impossible to know what uses would have been found for these data. Since it is impossible to know what uses would have been found for these data.

Moreover, the types of data being collected change through time. We attempted to control for these effects by focusing on a single type of data that has been collected in the same way for many decades: data on morphometric data.

It is likely that expectations on data sharing will differ between academic communities and that some data types are easier to preserve than others. It is hard to tell whether this difference is due to the slightly different research profiles such as ResearchGate or Google Scholar that asking authors for their data shortly after publication from papers that were less than a year old.

Considering only the papers from 2011, our results show that researchers any chance of reusing them. Fortunately, one important, leaving their preservation to authors denies future researchers any chance of reusing them. Fortunately, one effective solution is to require that authors share it on a public archive at publication: the data will be preserved in perpetuity. Researchers any chance of reusing them.

ResearchGate or Google Scholar that asking authors for their data shortly after publication from papers that were less than a year old. It is hard to tell whether this difference is due to the slightly different research profiles such as ResearchGate or Google Scholar that asking authors for their data shortly after publication from papers that were less than a year old. It is hard to tell whether this difference is due to the slightly different research profiles such as ResearchGate or Google Scholar that asking authors for their data shortly after publication from papers that were less than a year old.

Considering only the papers from 2011, our results show that researchers any chance of reusing them. Fortunately, one important, leaving their preservation to authors denies future researchers any chance of reusing them. Fortunately, one effective solution is to require that authors share it on a public archive at publication: the data will be preserved in perpetuity. Researchers any chance of reusing them.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages. The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent's attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project. The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent’s attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those published after 2000.

### Table 1. Breakdown of Data Availability by Year of Publication

<table>
<thead>
<tr>
<th>Year</th>
<th>No Working E-Mail</th>
<th>No Response to E-Mail</th>
<th>Response Did Not Give Status of Data</th>
<th>Data Lost</th>
<th>Data Exist, Unwilling to Share</th>
<th>Data Received</th>
<th>Data Extant (Unwilling to Share + Received)</th>
<th>Number of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>9 (35%)</td>
<td>9 (35%)</td>
<td>2 (8%)</td>
<td>4 (15%)</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>1993</td>
<td>14 (39%)</td>
<td>11 (31%)</td>
<td>3 (8%)</td>
<td>7 (19%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td>1 (3%)</td>
<td>36</td>
</tr>
<tr>
<td>1995</td>
<td>11 (31%)</td>
<td>9 (26%)</td>
<td>0 (0%)</td>
<td>7 (20%)</td>
<td>2 (6%)</td>
<td>6 (17%)</td>
<td>8 (23%)</td>
<td>35</td>
</tr>
<tr>
<td>1997</td>
<td>11 (37%)</td>
<td>9 (30%)</td>
<td>1 (3%)</td>
<td>2 (7%)</td>
<td>3 (10%)</td>
<td>4 (13%)</td>
<td>7 (23%)</td>
<td>30</td>
</tr>
<tr>
<td>1999</td>
<td>19 (48%)</td>
<td>13 (32%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>6 (15%)</td>
<td>6 (15%)</td>
<td>40</td>
</tr>
<tr>
<td>2001</td>
<td>13 (30%)</td>
<td>15 (35%)</td>
<td>3 (7%)</td>
<td>4 (9%)</td>
<td>0 (0%)</td>
<td>8 (19%)</td>
<td>8 (19%)</td>
<td>43</td>
</tr>
<tr>
<td>2003</td>
<td>9 (20%)</td>
<td>20 (43%)</td>
<td>4 (9%)</td>
<td>2 (4%)</td>
<td>0 (0%)</td>
<td>11 (24%)</td>
<td>11 (24%)</td>
<td>46</td>
</tr>
<tr>
<td>2005</td>
<td>11 (24%)</td>
<td>14 (31%)</td>
<td>6 (13%)</td>
<td>1 (2%)</td>
<td>0 (0%)</td>
<td>13 (29%)</td>
<td>13 (29%)</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>12 (18%)</td>
<td>31 (47%)</td>
<td>2 (3%)</td>
<td>4 (6%)</td>
<td>1 (2%)</td>
<td>16 (24%)</td>
<td>17 (26%)</td>
<td>66</td>
</tr>
<tr>
<td>2009</td>
<td>9 (13%)</td>
<td>34 (49%)</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>6 (9%)</td>
<td>12 (17%)</td>
<td>18 (26%)</td>
<td>69</td>
</tr>
<tr>
<td>2011</td>
<td>13 (16%)</td>
<td>29 (36%)</td>
<td>8 (10%)</td>
<td>0 (0%)</td>
<td>7 (9%)</td>
<td>23 (29%)</td>
<td>30 (38%)</td>
<td>80</td>
</tr>
<tr>
<td>Totals</td>
<td>131 (25%)</td>
<td>194 (38%)</td>
<td>33 (6%)</td>
<td>37 (7%)</td>
<td>20 (4%)</td>
<td>101 (19%)</td>
<td>121 (23%)</td>
<td>516</td>
</tr>
</tbody>
</table>

Data are displayed as n (%); the percentages are calculated by rows.
We found a strong effect of article age on the availability of data from these 516 studies. The decline in data availability could arise because the authors of older papers were less likely to respond, but this was not supported by the data. Instead, researchers were equally likely to respond (Figure 1B) and to indicate the status of their data (Figure 1C) across the entire range of article ages.

The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media. For papers where authors reported the status of their data, the odds of the data being extant decreased by 17% per year (Figure 1D). There was a continuum of author responses between the data being reported lost and being stored on inaccessible media, and they seemed to vary with the amount of time and effort involved in retrieving the data. Responses included authors being sure that the data were lost (e.g., on a stolen computer) or thinking that they might be stored in some distant location (e.g., their parent’s attic) to authors having some degree of certainty that the data are on a Zip or floppy disk in their possession but no longer having the appropriate hardware to access it. In the latter two cases, the authors would have to devote hours or days to retrieving the data. Our reason for needing the data (a reproducibility study) was not especially compelling for authors, and we may have received more of these inaccessible data sets if we had offered authorship on the subsequent paper or said that the data were needed for an important medical or conservation project.

The odds that we were able to find an apparently working e-mail address (either in the paper or by searching online) for any of the contacted authors did decrease by about 7% per year. This decrease was partly driven by a dearth of e-mail addresses in articles published before 2000 (0.38 per paper on average for 1991–1999) compared with those after 2000.
Public Data Archiving: Who Shares? Who Doesn't? What Can We Do About It?

H. Piwowar