Alyssa Goodman

Edit this form

173 responses

Summary See complete responses

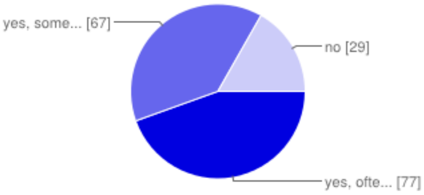
A little about you...

In what year did, or do you expect to, receive a PhD?

2008 2012 1989 2013 2012 2010 1989 1983 1985 2007 2014 2014 1986 1996 1976 2017 1983 2010 2010 1996 1968 1991 2007 2016 20

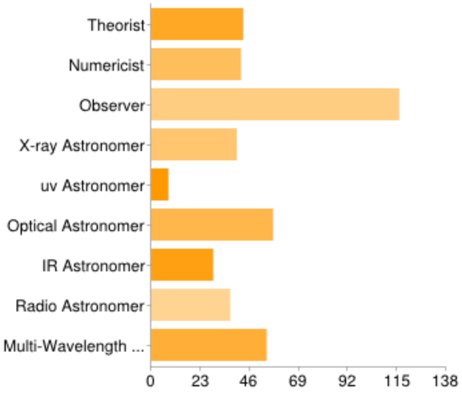
Your DATA-related Story

Do you use data from large (e.g. NASA) archives?



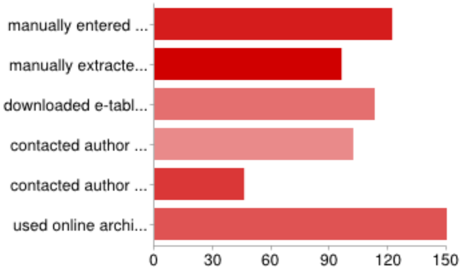
yes, often (once a month or more)	**77**	45%
yes, sometimes (less than once a month)	**67**	39%
no	**29**	17%

Do you consider yourself a...?



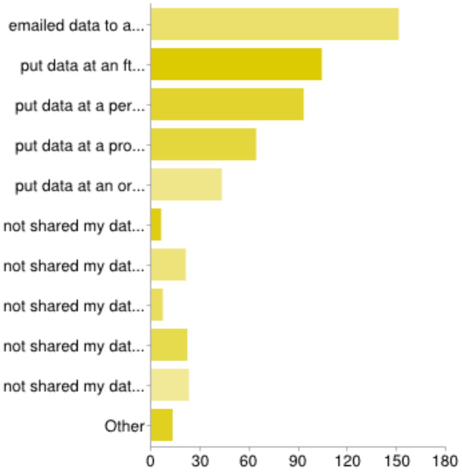
Theorist	**43**	10%
Numericist	**42**	10%
Observer	**116**	27%
X-ray Astronomer	**40**	9%
uv Astronomer	**8**	2%
Optical Astronomer	**57**	13%
IR Astronomer	**29**	7%
Radio Astronomer	**37**	9%
Multi-Wavelength Astronomer	**54**	13%

Have you ever used DATA you learned about from reading a Journal article?



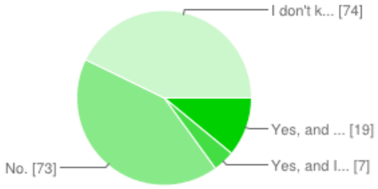
manually entered data from a table in a paper	**122**	19%
manually extracted data point vaues from a graph	**96**	15%
downloaded e-table of ASCII data provided by Journal	**113**	18%
contacted author to ask for data & got what I needed	**102**	16%
contacted author to ask for data & did NOT get what I needed	**46**	7%
used online archive where data were available	**150**	24%

When it comes to sharing DATA you've created, collected or curated, you have



emailed data to a colleague upon request.	15
put data at an ftp-style site for a colleague to retrieve.	10
put data at a personal web site, such as http://cfa-www.harvard.edu/~agoodman	9
put data at a project-based web site, such as http://www.cfa.harvard.edu/COMPLETE/data_html_pages/data.html	6
put data at an organized institutional archive, such as http://theastrodata.org or http://tdc-www.harvard.edu	4
not shared my data, because I think it will endanger my career.	
not shared my data due to large file sizes	2
not shared my data because I don't know how.	
not shared my data because it takes too much effort.	2
not shared my data because I don't think anyone will want it.	2
Other	1

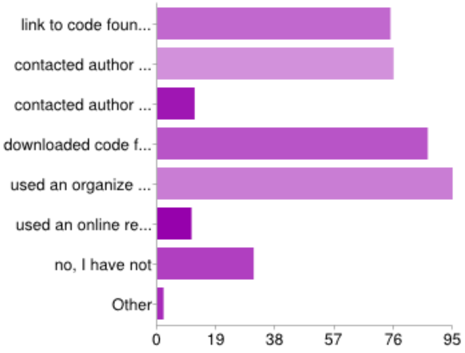
I am subject to NSF data management plans



Yes, and I have solutions.	**19**	11%
Yes, and I need solutions.	**7**	4%
No.	**73**	42%
I don't know what that means.	**74**	43%

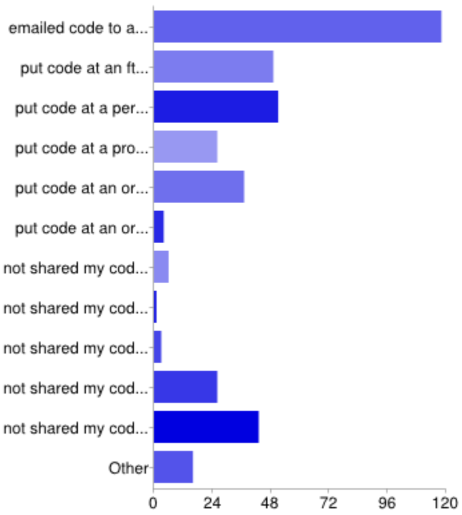
Your CODE-related Story

Have you ever used CODE you learned about from reading a Journal article?



link to code found in a Journal paper	**75**
contacted author to ask for code & got what I needed	**76**
contacted author to ask for code & did NOT get what I needed	**12**
downloaded code from non-repository web-site (e.g http://projects.iq.harvard.edu/seamlessastronomy/software).	**87**
used an organize code repository, such as github.com, sourceforge.net, or asci.net	**95**
used an online repository not intended specifically for code, such as theastrodata.org	**11**
no, I have not	**31**
Other	**2**

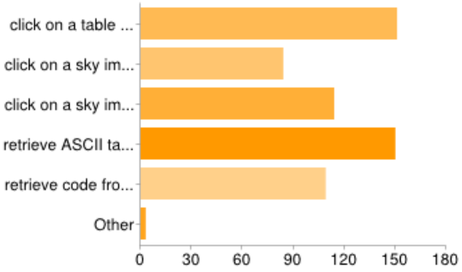
When it comes to sharing CODE you've created, collected or curated, you have



emailed code to a colleague upon request.	**1**
put code at an ftp-style site for a colleague to retrieve.	
put code at a personal web site, such as http://cfa-www.harvard.edu/~agoodman	
put code at a project-based web site, such as http://www.cfa.harvard.edu/COMPLETE/data_html_pages/data.html	
put code at an organized code repository such as github.com, sourceforge.net, or asci.net	
put code at an organized repository not intended specifically for code, such as theastrodata.org	
not shared my code, because I think it will endanger my career.	
not shared my code due to large file sizes	
not shared my code because I don't know how.	
not shared my code because it takes too much effort.	
not shared my code because I don't think anyone will want it.	
Other	

About the Future

If tighter data-code-literature connections like those listed below were available, I would often (more than once a year):



click on a table in a PDF or HTML view and be offered an ASCII download

click on a sky image in a PDF or HTML view and be offered to see that image as an overlay on a selectable-wavelength

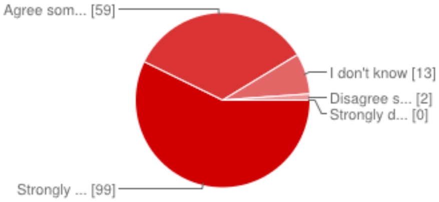
click on a sky image and retrieve a FITS file for that image

retrieve ASCII table versions of information shown in graphs

retrieve code from URLs offered within a paper

Other

I think that the future of astrophysical research will rely more on sharing of code and data in the future than it has in the past.



Strongly agree	**99**	57%
Agree somewhat	**59**	34%
I don't know	**13**	8%
Disagree somewhat	**2**	1%
Strongly disagree	**0**	0%

A little more about you (fully optional!)

Additional Comments (optional)

Ideally, the version of code employed in papers can be embedded in a virtual machine (VM) and included together with the publication. This not only makes verifying and validating results easier, but saves countless hours spent running configure, make, make install and swearing at gcc and g77/gfortran compilers. The survey specifies ASCII, but realistically given the sizes of current data sets, HDF5/FITS binary tables are perfectly fine. Very large data sets need wrappers to generate SQL or some other database dump. You can't be using grep/awk/sed/find/numpy.where with even a couple of GB of data. Major code libraries (numpy/scipyAstroPy, IDL Astronomy Library, CFITSIO, IRAF...) really need modification to be aware of numerical representations, units and dimensions. Too many errors are the result of accidentally misusing data (passband transmissions reported in ergs/Angstrom in synthetic photometry equations derived with passbands represented in dimensionless photon efficiency being my particular pet peeve). This is a particular issue with multi-wavelength studies and should be transparent - i.e. something the end user shouldn't actually have to do manually other than specifying the output format they'd like. Combining someone's F_lambda optical spectrum in Angstroms and ergs cm^-2 s^-1 Angstrom^-1 with F_nu radio data reported in Janskies and GHz for example. Another is any table that has a column labelled redshift without specifying if it's heliocentric, CMB frame.... that is again, purely an issue of representation of the data and could be addressed easily by metadata in a binary table, together with libraries that were actually aware of that metadata. Experimental physicist. Your checkboxes did not admit of that answer, and would not allow me to exit the survey without specifying one of the answers I feel like I'm not too bad at data sharing (and believe strongly in it, particularly when the funding is from the taxpayer), but find when it comes to code sharing my instinct is that my code is too messy / sloppy and not general enough for outside use, and it would take too much time to clean it up and document it properly, and then possibly have to support it. Another way to say this is it doesn't meet my self-imposed standards for publication as-is. While I always try to provide it on request, such requests are (from experience) very rare, and I think the only ones I have received were from ex- or current collaborators. I feel like this is a common problem. It works the other way around too, I have been known to spend days working on transcribing data tables by hand, or reimplementing a numerical method myself, rather than trying to contact the author to see if they still have a machine readable version. One general problem in my particular area of interest (low-mass stars in the solar neighborhood) is that publication of new data is incremental, and there are limited resources compiling all of the available data for a given object. Doing this as and when needed (e.g. for a paper) can be extremely time consuming and prohibits many useful statistical studies without putting vast efforts into compiling data from the literature. The point here is what we currently seem to lack as a community is a good way to link the various datasets together to build a body of knowledge that is not just a few hundred separate papers with inconsistent identifiers, table formats, etc. scattered around the literature. The particular issue with the solar neighborhood that unfortunately limits use of some of the excellent resources out there for doing catalog cross-matching is that this cannot be done on position alone due to the high proper motions and multiple nature of many of the sources, which cause a great deal of matching errors. I have been working for a while on my own solution to producing many of the compilations I need in my research, and hope to be able to make all of this work available to the public if enough time can be found to finish it. There are a large number of scientists studying solar and heliospheric physics at CfA. The NASA Heliospheric Division data policy is that all data are open as soon as the calibrations have been applied (typically 1-2 weeks). This is a very different environment from the Astrophysics division proprietary data. The main difficulty is that the volume of data is large by most archive standards (2.5TB/day from SDO). Papers typically use a small subset of that data, but the trend in papers is to analyze statistically significant samples of observations. Storing the full data sets associated with papers may be prohibitively expensive. Of course numerical simulations now have the same problem or worse. If you want to reproduce the results of 3-D time dependent simulation that is going to require the saved solutions from the simulation. I don't think it is reasonable to expect a scientist to run someone else's massive 3D code any more than you can expect someone to run a telescope. In my experience, state of the art codes are fragile and require expert handling. there is always the issue of folks not wanting to share too much if they have invested a lot - where others could take advantage. So, code and data sharing seem to make the most sense in the context where it is "mature" data and code. All code and data should (eventually) be publicly available. Restricting access to code or data is contrary to the very nature of not only the scientific process, but the entire goal of scientific study. For observational data, storing proper metadata (before, during, and after the observing run) is a cultural shift, made more complex by the growth of multi-object instruments. It is something that many astronomers resist, primarily because they think it will take more time, but sometimes because they do not want to release "their" data on any useful timescale. I suspect that it will require *institutional* pressure to get it to happen routinely; i.e., being a requirement to use the telescopes. It would be great if all data from a paper was available directly from the journal or some kind of link in the paper (that wouldn't expire or go bad). The journals should do more to enable/encourage/require it from authors. Having it on personal web sites or even project or NASA sites is not as good because those links will eventually break or get moved. Just a bit on the history as I see it for making astronomical data publicly available. For the Einstein observatory, Riccardo Giacconi, who was the Einstein PI, led the effort to make the data public. He left CfA to become STScI director and of course there is a public archive of HST data. Riccardo then became Director General of ESO and the VLT data is publicly available. Today we expect that data from NASA missions will of course be publicly available in one year or less. I think Riccardo had a lot to do with this. Resources have always been precious and access to observatories is limited. Making data as well as the tools to analyze them (including ones like ds9) widely available is good for science. I think it's a crucial step (code and data sharing) that is a severe deficiency in our current process. More standardized ways to link versioned code and data with publications is something the community needs! I am a solar physicist. The vast majority of data in my field is freely open to everyone, so that probably colors my opinions somewhat. I think it would be very important to be able not only to share data, but to do so in an interactive way, so the user can access the different aspects of a multidimensional dataset (i.e. wavelength/energy, position, velocity, etc.), not only from observations, but also from large simulations. Gamification will/should be the name of the game! I'm not really an astronomer-- my answers above relate to spectroscopic and geophysical data and code. I like the idea of data points in plots being (always) available. I don't know if this is pertinent here, but I believe that the scientific community should make efforts towards sharing all the information necessary to reproduce a given scientific publication. Authors should be encourage to provide, along with their results, the reduced observations they have used, models applied, and so on. I would even include the data and, if available, script code used to reproduce each figure in the paper. That is, in my opinion, what the community should be aiming for in terms of data and code sharing. I have primarily relied on code supplied by our central servers like IRAF AIPS My answer to the question: "Have you ever used CODE you learned about from reading a Journal article?" is: No. However, to complete the survey, I was forced to click at least one answer. Therefore, please ignore my answer to that question, it is made up. Despite my lack of a PhD, I am a full member of the AAS and IAU, have published scientific papers, and shared both data and code. I'm on the advisory committee for the Astrophysical Source Code Archive and was invited to the Library of Congress for a meeting on code archiving. I've distributed my own software and catalogs on the web for almost 20 years and via tapes and ftp before that Part of my experience is based on a large radiative transfer code for planet atmospheres that I wrote many years ago, which has evolved over time, and was given to Lisa Kaltenegger for our joint use, now effectively her use only, when she came to work with me at the CfA. The other main part of my experience is with the very large database being generated daily by the Kepler mission, which I use daily in order to extract statistical results, for example intrinsic biases and a prediction of the frequency of terrestrial planets in habitable zones, and here I have learned a lot about how to handle files with a lot of data on 150,000 stars, and for which it would be useful to have code sharing, since there is none in this area, and I need to write all my analysis codes from scratch. I like the idea of data points in plots being (always) available. What about "Do you consider yourself a... Developer"? (or something equivalent) There are a lot of observations which do not get published for various reasons e.g. insufficient data within the group, long delay in drafting and publishing and ultimately the idea of writing paper gets discarded etc. That data could, however, be of significant importance and can be used far more efficiently if made available to the community. But such groups probably need assistance in putting the data together and out without which the data and the large amount of resources spent on acquiring it in the first place will be wasted. Despite not having ever denied code or data on request because I felt sharing could harm my career, I think there are situations in which that could be the case. This survey may not take into account our particular database and codes. We develop, manage, and distribute molecular spectroscopic data. One of the primary uses of the database is to serve as input to radiative-transfer codes. Although there are many benefits to code and data sharing, I think it shouldn't be made TOO easy. Not to actively discourage people, but to make sure that people do the required leg-work to really understand what they're getting. I'm all for sharing tabular and FITS data, but the whole "click on an image and get all sorts of prizes" thing seems like it absolves the user of having to know what they're doing. github is really great for sharing code. It is infinitely better than e.g. trying to e-mail code to people or trying to give them passwords to access versioning repositories on Harvard-specific resources. I wish github would allow me to restrict who can/cannot access my repositories without requiring me to be a paying member. Or maybe this situation has changed since the last time I investigated... It would be great if ADS MRTs included all of the data that was already available electronically through the journal. I've made online-only versions of tables for ApJ, but not seen them on ADS. I *hope* that the future will involve more sharing of code and data, but I worry about the incentive structure of the field. It takes effort to prepare code and data for circulation, and often the reward for that effort is unclear. Code sharing (last question above) is often more of a pain than an asset (code not transportable; not exactly applicable to different needs; etc.). So am not convinced code sharing is the answer (but in some cases surely useful. We are about to announce (c. May 1) the first data release for DASCH (DR1). This will all be via a new DASCH website. The photometry and astrometry code that produced ~3 billion mag. measures (yes, that's what 9% of the plates now scanned result in!) are not being put "out there" since its a MASSIVE processing pipeline and simply not easily transported or separable. Maybe that's the reason for some of my doubts about utility of "code sharing".

Your First Name (optional!)

Xavier jeremy Gautham Michael Jim Jonathan Xuejian Edward John Kate Randall Willie Luke Zach Christine Cameron Christopher Lauranne Henry
Sen Richard David Sven Atish Matt Til Ewan John Laurence Bruna Christopher Yuan-
Sen Michael Howard Ken Brian Aaron Courtney Paul Katrien Moritz Ting Scott Jaaon Matt John Peter Jonathan Alexey Raffaele Colin Josh

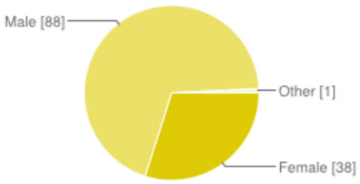
Your Last Name (optional!)

Dumusque drake Narayan McCollough Phillips Irwin Jiang DeLuca Raymond Alexander Smith Torres Kelley Berta Jones McBride Stubbs Lanz Winter
Galarza Nelson tucker Qi Garcia Copete Laskar Seward Mink Traub Nelson Czekala Allen Ting Edgar Phillips Van
Loo Kamble Holman Birnstiel O'Sullivan Ruan Rothman Vajgel Faesi Ting Stevens Smith Young Stalder Meisner Dressing Torrey Kolenberg guenther

Your email address

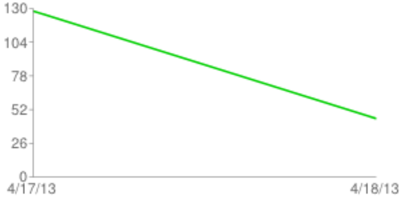
xdumusque@cfa.harvard.edu jdrake@cfa.harvard.edu gnarayan@cfa.harvard.edu lingzhen@cfa.harvard.edu mmccollough@head.cfa.harvard.edu jphillips@cfa.harvard.edu
sen.ting@cfa.harvard.edu inicica@yahoo.com redgar@cfa.harvard.edu dphil@cfa.harvard.edu svanloo@cfa.harvard.edu bbarsdell@cfa.harvard.edu matthewjohnpayne@g
sen.ting@cfa.harvard.edu mstevens@cfa.harvard.edu kyoung@cfa.harvard.edu brian.stalder@gmail.com fsienkiewicz@cfa.harvard.edu ameisner@fas.harvard.edu cdressi

Gender (optional!)



Female	**38**	30%
Male	**88**	69%
Other	**1**	1%

Number of daily responses


130
104
78
52
26
0
4/17/13
4/18/13