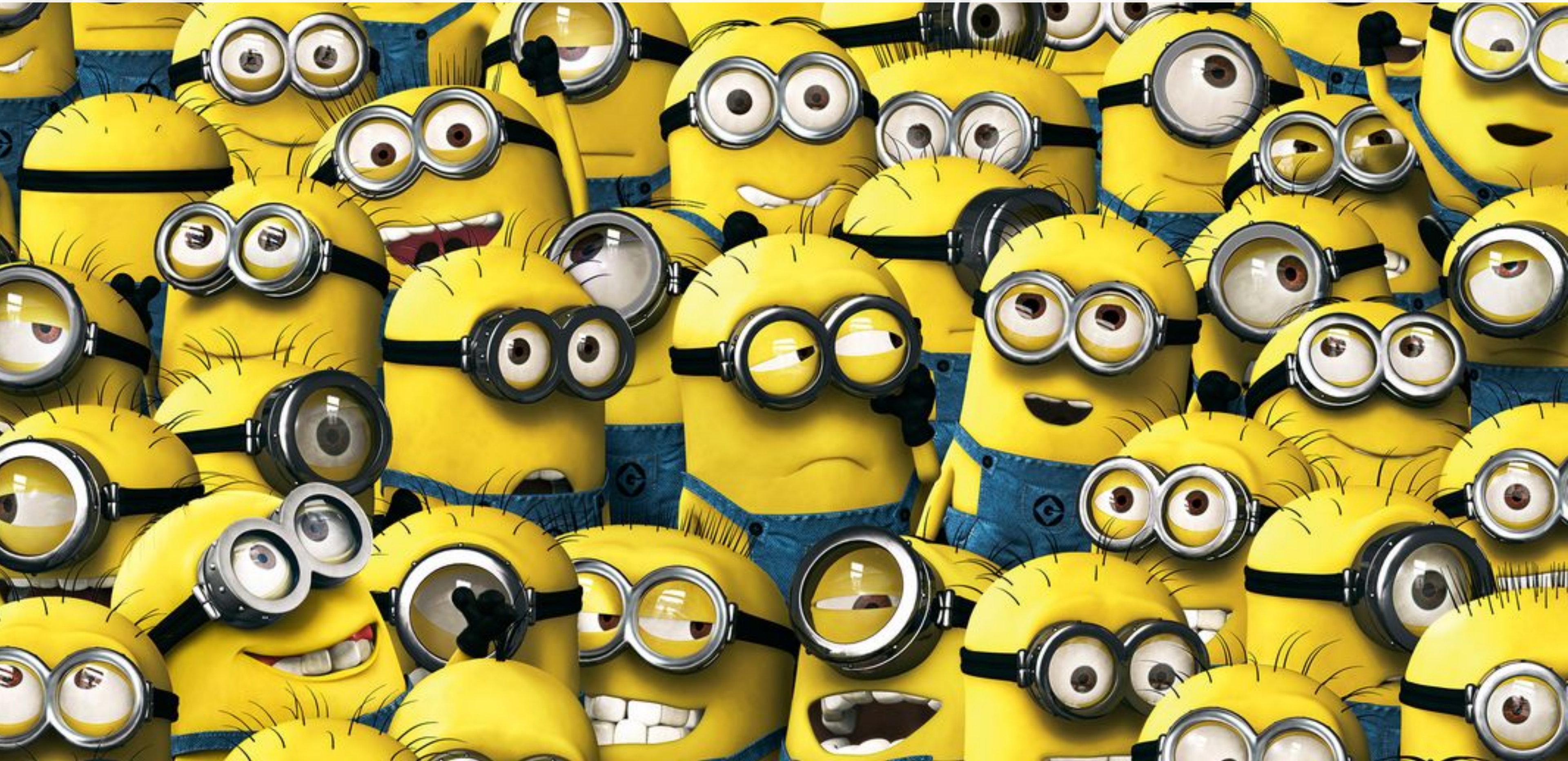
If you had 100,000 people to help you with your research, what would you do?



What is Open Science?

- Open Access to articles and lab notebooks
- Open Data
- Open Source Code
- Open Collaboration (e.g., citizen science)
- Open Technology (e.g., Makers)
- Open Funding

OECD (2015), "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en

Publishing Your Code Open Source: Science Code Manifesto

- **Code:** All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.
- **Copyright:** The copyright ownership and license of any released source code must be clearly stated.
- **Citation:** Researchers who use or adapt science source code in their research must credit the code’s creators in resulting publications.
- **Credit:** Software contributions must be included in systems of scientific assessment, credit, and recognition.
- **Curation:** Source code must remain available, linked to related materials, for the useful lifetime of the publication. “

Source: Sciencecodemanifesto.org

Example: Reproducible Research Workflow for Analysis of Human Microbiome Data

- “After obtaining amplicon sequence, a standard series of bioinformatic and statistical analysis are used to evaluate the data: filtering out low quality sequences and samples, constructing a taxonomic feature table of observations from each sample, incorporating the sample metadata, transforming and normalizing the feature table, and performing exploratory and inferential analyses. “

- Through a bioinformatic forensics of Arumugam et al (2011), published in Nature, authors estimated **“more than 200 million possible ways of analyzing these data.”** (p. 185)

- Methods used to filter sequences, construct taxonomic features, and perform analysis **“are often performed in separate environments**, **making creation of a single coherent record of the analysis difficult and time consuming.”** (p. 186)

Example: Reproducible Research Workflow for Analysis of Human Microbiome Data

- Researchers need **to "adopt pipelines documenting choices used in these analysis** with the intention of providing an assessment of the robustness and reproducibility of the analysis." (e.g., LaTex, Rmarkdown, JupyterNotebooks)

- Authors advocate platforms like GitHub to **deposit data, associated materials** (e.g., complete metadata mapping files, taxonomy files, reference sequence files), and **code.**

- **"Publish reproducible workflows encompassing the entirety of the analysis"** (e.g., unified R script and unified Rdata data object)

Callahan et al.. 2016. Reproducible Research Workflow in R for the Analysis of Personalized Human Microbiome Data, Pacific Symposium on Biocomputing 2016: https://www.ncbi.nlm.nih.gov/pubmed/26776185

Accepting Open Source Contributions

- Shift from low-value work to to high-value work
- Facilitates rapid prototyping and experimentation
- Advances interoperability between tools
- Lower total cost of ownership
- Detect, diagnose, triage, and resolve bugs faster
- Increases reliability & security through peer review
- Shift workflows from waterfall to agile and lock-free
- Encourage more modular code
- Reduce duplication of effort
- Attract talented developers

Sources: Ben Balter, GitHub, https://ben.balter.com/2015/11/23/why-open-source/
18F Open Source Policy https://github.com/18F/open-source-policy/blob/master/policy.md

Example: Google's TensorFlow

TensorFlow is a research-grade, open-source software library for dataflow programming across a range of tasks, including machine learning applications.

- Google is exposed to emerging academic research through the open source community.
- **Google engineers design the interface, test and validate code that is introduced, and sync the internal and external versions.**
- Beyond TensorFlow, Google uses a large number of open source libraries in tis production code, increasing the economic impact.
- **Cloud computing** for big data and runnable instances of code!

Source: David Konerding, Google Open Science team

TensorFlow Code of Conduct

"In the interest of fostering an open and welcoming environment, we as contributors and maintainers pledge to making participation in our project and our community a harassment-free experience for everyone, regardless of age, body size, disability, ethnicity, gender identity and expression, level of experience, nationality, personal appearance, race, religion, or sexual identity and orientation."

Source: https://github.com/tensorflow/tensorflow/blob/master/CODE_OF_CONDUCT.md

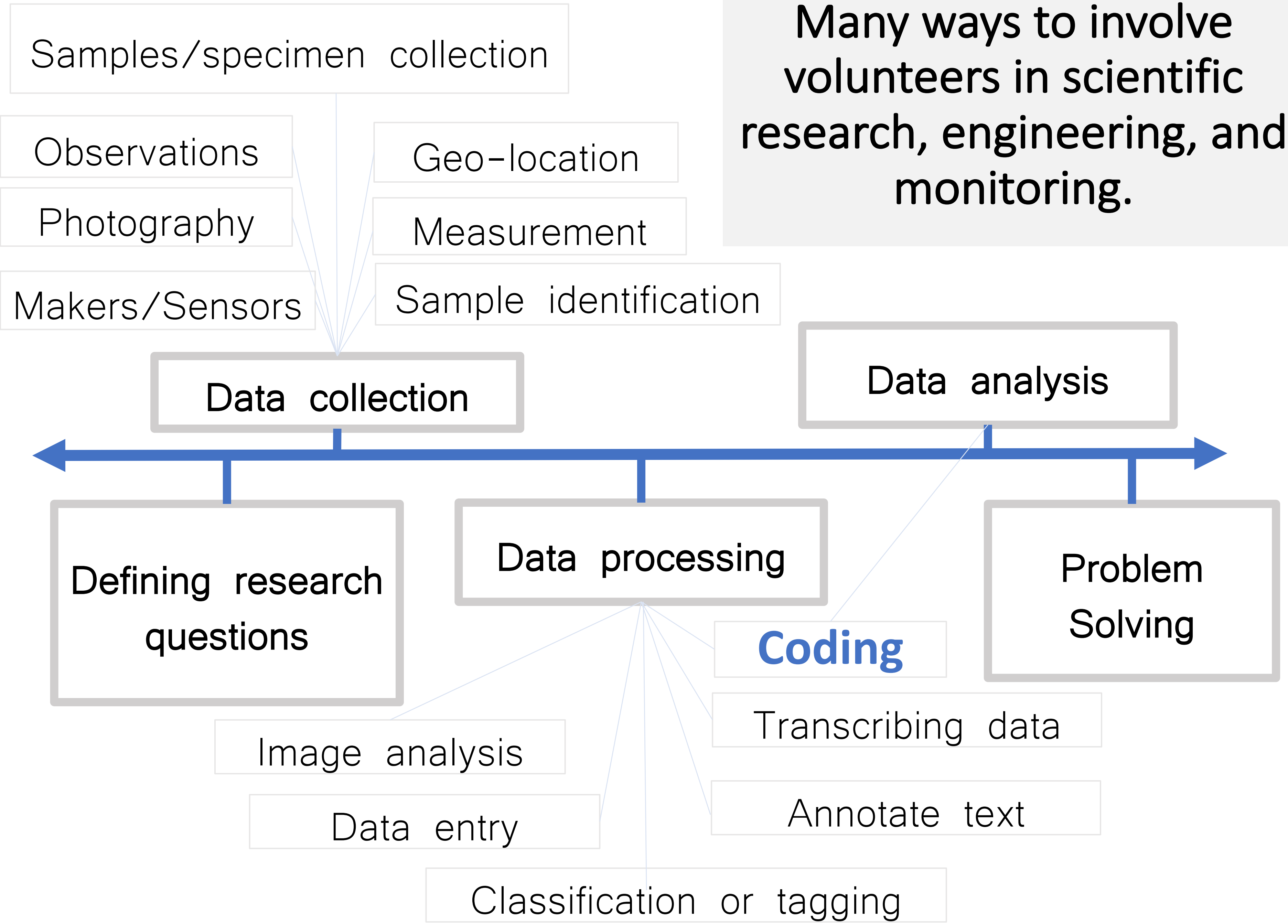
"Project maintainers are responsible for clarifying standards of acceptable behavior and are expected to take appropriate and fair corrective action" as necessary.

What is Citizen Science?

Contributions of the public to the advancement of scientific and engineering *research and monitoring* in ways that may include:

- Identifying research questions
- Designing/conducting investigations
- Designing/building/testing low cost sensors
- Collecting and analyzing data
- Developing data applications
- Developing technologies for science
- Solving complex problems

Many ways to involve volunteers in scientific research, engineering, and monitoring.



Samples/specimen collection

Observations

Geo-location

Photography

Measurement

Makers/Sensors

Sample identification

Data collection

Data analysis

Defining research questions

Data processing

Problem Solving

Coding

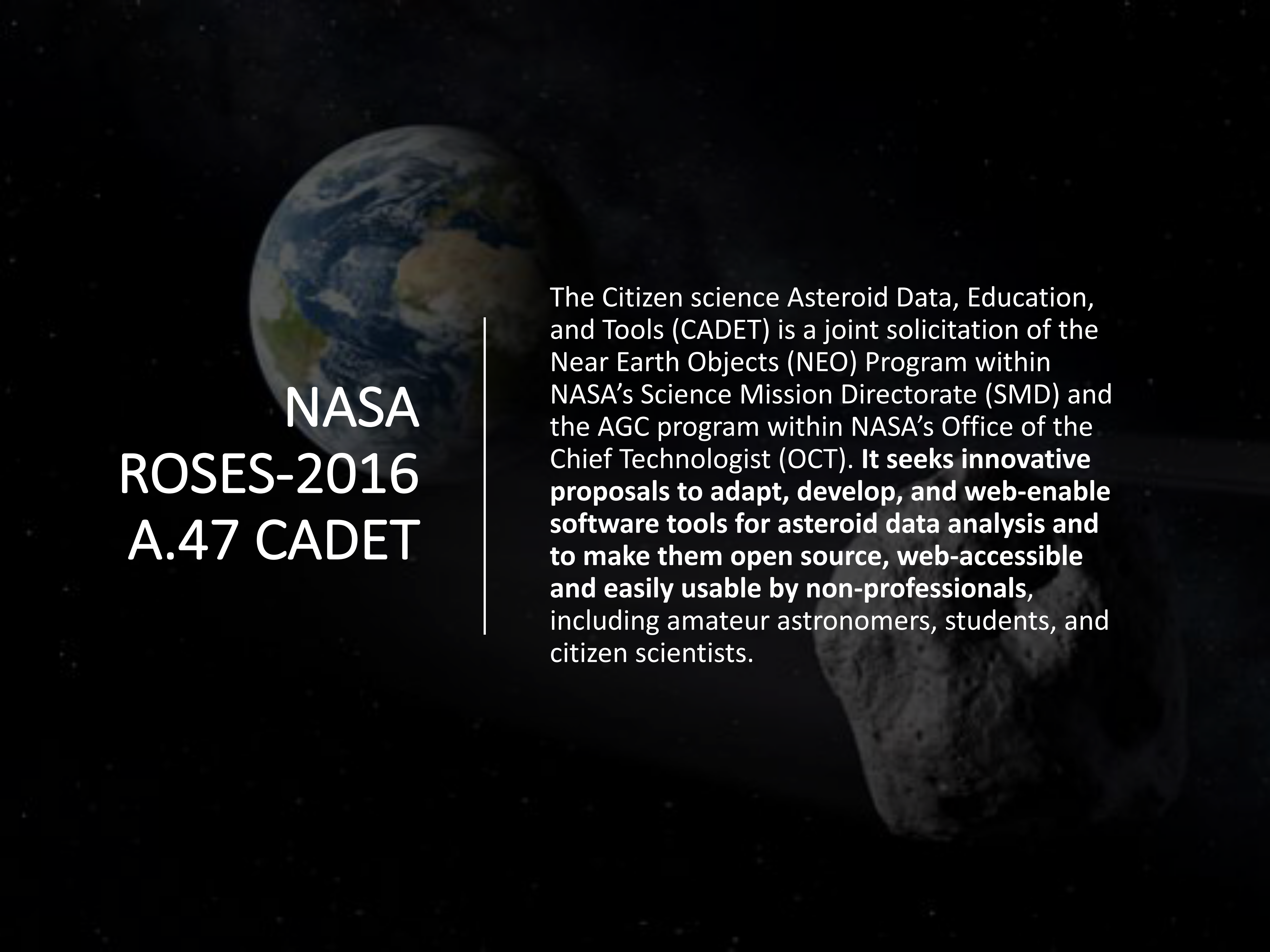
Transcribing data

Image analysis

Data entry

Annotate text

Classification or tagging



NASA ROSES-2016 A.47 CADET

The Citizen science Asteroid Data, Education, and Tools (CADET) is a joint solicitation of the Near Earth Objects (NEO) Program within NASA’s Science Mission Directorate (SMD) and the AGC program within NASA’s Office of the Chief Technologist (OCT). **It seeks innovative proposals to adapt, develop, and web-enable software tools for asteroid data analysis and to make them open source, web-accessible and easily usable by non-professionals**, including amateur astronomers, students, and citizen scientists.

NASA ROSES-2016 A.47 CADET

The specific goals of the CADET program are to:

- Through **agile development** and other innovative methods, adapt, further develop and web-enable asteroid data analysis software to increase the productivity of NEO and AGC research endeavors and extend the state-of-the practice in those endeavors;
- Develop **easily usable and understandable software tools though the application of human-centered design best practices**, including user research studies, systematic usability testing, and evaluation;
- Integrate advances in information technology with advances in cyberlearning (i.e., what is known about how people learn with technology), and integrate these software tools into learning environments so their potential is fulfilled;
- **Foster multi-disciplinary collaborations that span the NASA science, computer science, design, and education disciplines.**

NASA ROSES-2016 A.47 CADET

- Astrometry: Solve an image for the positions of the stars and moving objects within it. Allow batch processing of images.

- Photometry: Provide relative photometry for the moving objects within an image using reference stars from standard catalogs. Allow for batch processing of images.

- Light Curve Analysis: Provide a virtual workbench for the user to create asteroid light curves from the derived photometric values and determine the spin period of the asteroid.
  - **The software must account for light-time correction and the phase angle effect on brightness**.
  - The software must produce…an image of the folded light curve that is suitable for publication in a scientific journal.

NASA ROSES-2016 A.47 CADET

- Because of the rapidly changing computing and computation technology environment, awards resulting from this call are are **limited to a performance period of 24 months or less.**
- Proposal must define **clear, measureable milestones.**
- The tools ….are expected to be available for open use as **web-based applications** and as **open source code.**
- **Plans for migration into a persistent software framework or infrastructure for ongoing maintenance and user support should be identified.** Proposals should discuss plans for this, even if the project is not expected to reach that level of maturity in the term of the award.
- Investigators are required to be archived the source code, and all relevant documentation, at **NASA's Github site** https://github.com/nasa

NASA ROSES-2016 A.47 CADET

- **Agile:** Proposals must include an agile software development and testing plan to describe the software engineering practice to be used in the project.
- **Human Centered Design:** Proposals also must address how they will develop and document user personas (e.g., high school and college students, amateur astronomers, and professional astronomers), engage end users in the iterative testing and evaluation of the software tool, and **how they will meet staffing expertise, as appropriate, including user experience (UX) design, user interface (UI) development, and web application development**. The proposal will be considered unresponsive without this plan.

NASA ROSES-2016 A.47 CADET

~~The use of Apache License 2.0 is required.~~ Any proposal responding to this solicitation to adapt, develop, and web-enable the asteroid data analysis software is required to deliver to NASA a copy of such software with sufficient rights for use as Open Source Software under ~~this Apache License 2.0.~~ Therefore, each proposal shall:

- Identify any proprietary software, software owned by a non-Federal entity, or open source software that is incorporated into the software being proposed;
- Indicate whether a license has been obtained in situations where proprietary software, software owned by a non-Federal entity, or open source software has been incorporated into the software that is the subject of the proposal and attach a copy of the license to the proposal, along with evidence of permission obtained from the software owner to release improvements or derivative works to the software as Open Source under ~~the Apache License, Version 2.0.~~ NASA will evaluate proposals for compliance with the above open source software requirements.
- A proposal that does not include documentation sufficient to satisfy NASA that the developed software will be open source may not be selected.

NSF Open Source Licensing Grant Requirements

- The NSF Future Internet Architecture-Next Phase (FIA-NP) program required open source licensing pursuant to the Open Source Initiative: https://www.nsf.gov/pubs/2013/nsf13538/nsf13538.htm.

- The NSF Secure and Trustworthy Cyberspace (SaTC) program states that PIs who choose to open source their software should employ a license listed by the Open Source Initiative: https://www.nsf.gov/pubs/2017/nsf17576/nsf17576.htm.

- **"Any software developed as part of this program is required to be released under an open source license listed by the Open Source Initiative (http://www.opensource.org/)."**

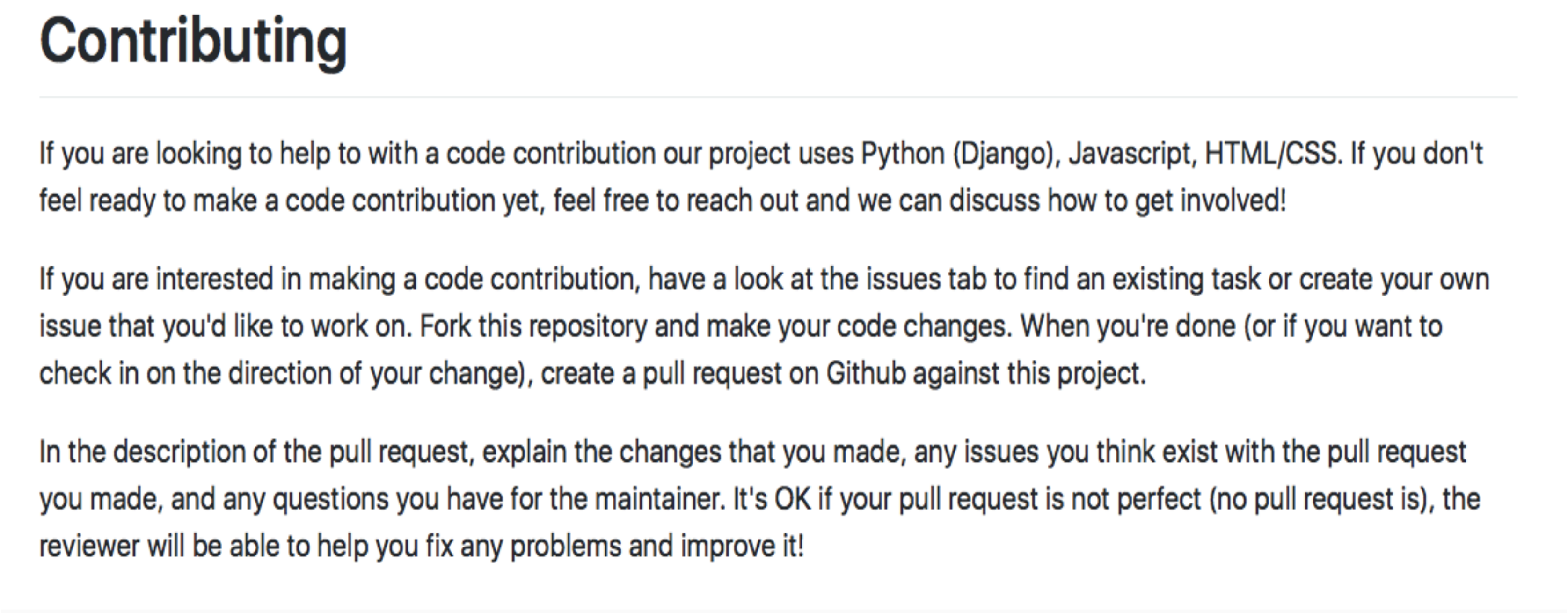
Open Source Management

- **Update and expand NASA's open source directory**
- Upload data & source code (e.g., NASA's Github, Bitbucket)
- Manage and publish reproducible workflows encompassing entire analysis (e.g., R, DI2E Jira)
- Provide templates for reuse of code
- Provide open documentation (e.g., bibliography of code library)
- **Don't bake passwords and sensitive info into code in public repo.**

Open Source Management

- **“Free like a puppy!”: Build and nurture the open source community through code-a-thons, online forums, and regular engagement activities (e.g., Slack, Gitter, Discord)**
- Provide guidelines for contributions (e.g., pull request policy).
- Establish code of conduct
- Encourage ”peer programming” or “mob programming”
- Make code modular and clean up as you go
- Provide open issue tracking (e.g., bugs, feature requests)
- Provide open road mapping (workflows, timelines)

Provide Guidance



Contributing

If you are looking to help to with a code contribution our project uses Python (Django), Javascript, HTML/CSS. If you don't feel ready to make a code contribution yet, feel free to reach out and we can discuss how to get involved!

If you are interested in making a code contribution, have a look at the issues tab to find an existing task or create your own issue that you'd like to work on. Fork this repository and make your code changes. When you're done (or if you want to check in on the direction of your change), create a pull request on Github against this project.

In the description of the pull request, explain the changes that you made, any issues you think exist with the pull request you made, and any questions you have for the maintainer. It's OK if your pull request is not perfect (no pull request is), the reviewer will be able to help you fix any problems and improve it!

Source: Astrokit.org and https://github.com/typpo/astrokit/blob/master/README.md

Additional Examples

- **HydroShare**: NSF Collaborative Research: SI2-SSI: An Interactive Software Infrastructure for Sustaining Collaborative Community Innovation in the Hydrologic Scienceshttps://www.nsf.gov/awardsearch/showAward?AWD_ID=1148453 and https://www.hydroshare.org (open source)
- **SciDas:** NSF CC*Data: National Cyberinfrastructure for Scientific Data Analysis at Scale: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1659300 and http://renci.org/research/scientific-data-analysis-at-scale-scidas/ (open source)
- iRods Open Source Data Management Software https://irods.org

Additional Examples

- **NIH Data Commons** is to accelerate new biomedical discoveries by providing a cloud-based platform where investigators can store, share, access, and compute on digital objects (data, software, etc.) generated from biomedical research and perform novel scientific research including hypothesis generation, discovery, and validation. It is expected that awardees will participate collectively as a consortium and work cooperatively toward achieving NIH's comprehensive vision for an interoperable, FAIR (Findable, Accessible, Interoperable and Reusable) compliant, multi-cloud NIH Data Commons founded on open source and open standards. The Commons will be designed to comply with the principles of making digital objects FAIR. https://commonfund.nih.gov/bd2k/commons and https://commonfund.nih.gov/sites/default/files/RM-17-026_CommonsPilotPhase.pdf

Additional Resources

- **Consumer Financial Protection Bureau GitHub**: https://cfpb.github.io
- **DOD** Open Source Software FAQ: http://dodcio.defense.gov/Open-Source-Software-FAQ/#OSS_and_DoD_Policy
- **GSA 18F** posted an excellent roundup by Will Slack and Britta Gustafson that gathers citations and sources: https://18f.gsa.gov/2016/08/08/facts-about-publishing-open-source-code-in-government/
- GSA 18F also hosted a guest post by **DHS** that shows a concrete example of open source code leading to agencies saving on work, fixing each other's bugs, and generally just developing a great strategic collaboration: https://18f.gsa.gov/2017/01/06/open-source-collaboration-across-agencies-to-improve-https-deployment/

Additional Resources

- Open Source Licensing by Lawrence Rosen with foreward by Prof. Lessig: http://www.rosenlaw.com/oslbook.htm
- Code 2.0 by Lessig: http://codev2.cc/download+remix/Lessig-Codev2.pdf
- Cathedral and the Bazaar: http://www.catb.org/~esr/writings/cathedral-bazaar/
- Choosing an open source license: https://www.techrepublic.com/blog/web-designer/choosing-an-open-source-software-license-for-your-development-project/
- Choose an open source license (GitHub): https://choosealicense.com and https://choosealicense.com

Dr. Lea Shanley
co-Executive Director
NSF South Big Data Innovation Hub
LShanley@renci.org
@Lea_Shanley

The South Big Data Hub is funded in part through a grant from the NSF.

Thank you to: Pamela Gay, NASA CADET and COSMOQUEST PI; Ian Webster, NASA CADET PI Astrokit.Org /CTO, Zenysis Technologies; Sarah Allen, Sam Gensburg; Bob Balance; Ben Balter, GitHub; Ray Idaszak, RENCI; Jacqueline Kazil, PIF/PyLadies; Eric Mill, 18F; Jennifer Hammock, Smithsonian; Mike Byrne, CFPB; James Miller, FCC