The Next Decade of Scientific Python

Stéfan J. van der Walt, K. Jarrod Millman

University of California, Berkeley

March 2020

Grant Purpose

Brief Purpose Statement

Our objective is to prepare scientific Python for the next decade of data science. To this end, we will: (1) improve common engineering infrastructure, (2) better coordinate core projects, (3) write a community vetted strategic plan, and (4) help the community develop grant proposals.

Description of Project

With an extensive and high-quality ecosystem of libraries, scientific Python has emerged as the leading platform for data analysis. This ecosystem is sustained by independent volunteers with separate mailing lists, websites, roadmaps, documentation, engineering and packaging solutions, and governance structures. Unfortunately, this also means that there is a lack of coordination that results in duplicated effort, disorganized documentation, breakage upon new releases, unintended performance regressions, and user confusion. Moreover, we have no venue for developing a formal, shared vision of the future.

The developers of these projects are technically able; but they have little time to coordinate efforts within their own projects, let alone focus on strategies for bringing the entire ecosystem together. This project will provide support where it is deeply needed.

We will achieve our aims through: (1) working on the cross-cutting concerns of the ecosystem; (2) forming an advisory committee; and (3) hosting an annual three-day forum of community leaders. The advisory committee will guide our engineering work and help us recruit community leaders to the annual forum. We will use the annual forum to help us prepare a community vetted strategic plan as well as help the community develop grant proposals to fund future work.

Engineering Infrastructure

We will support engineering efforts across the ecosystem. To solidify scientific Python, we need the most commonly used packages to be well maintained. We will contribute time to engineering efforts on selected libraries, to assist with engineering needs and solution. Specific shared improvements include:

- **Shared Build Infrastructure and Binary Packaging** Each project currently maintains its own continuous integration testing configuration, and relies on a single volunteer to provide tooling for binary release packages.
- **Integrated Websites and Expanded Documentation** Users who switch from commercially driven packages often note that, while the Python ecosystem provides many tangible benefits, it lacks a feeling of unity. This is no small surprise: any solution combines many, independently developed libraries. We will improve coherence and accessibility by setting up a central web portal for scientific Python, that gives access to comprehensive documentation that spans the various projects. We will pay particular attention to numpydoc, the library used to parse and present function documentation for most scientific Python libraries.
- **Shared Benchmarking** To identify performance regressions, projects utilize Air Speed Velocity. There is no central location for hosting this infrastructure, and no standard patterns for deployment.
- **Documentation and Bug Days** Early on, we held several documentation and bug-fixing events that significantly reduced the number of outstanding issues and onboarded several important community members. Our purpose here will be to organize the event, announce it widely, and then coordinate with participants to ensure rapid feedback on and inclusion of their work.
- **OSS DevOps** Most projects in the ecosystem share similar workflows around GitHub, and while the platform has dramatically increased collaboration, that same surge also burdened developers with an increased amount of information to track. To simplify and streamline project management, we have we been developing tools at BIDS that make it easier to generate pull requests, search issues and pull requests offline, and generate project health reports.
- **Python 3 Modernization** Many projects have decided to drop Python 2 support at the end of 2019 or before. To support both versions of the language requires some scaffolding, which can

now be torn down. In addition, new functionality that can be used includes keyword-only arguments, a matrix-multiply operator, formatted strings, and an extended statistics module.

Project Coordination

- **Online community manual** Disseminate knowledge about good operating practices, such as how to build a welcoming community (viable governance structures, welcoming code of conduct, onboarding), generating binary wheels, supporting installation platforms (operating systems, conda/pip), through an online community manual.
- **Joint Governance Structures** Good governance is hard, and constructing documents such as Codes of Conduct can consume months of developer time. We will jointly develop sets of such documents that reflect the broadly aligned vision of the community, allowing adoption after only minimal adjustment.
- **Coordinated Release Cycles** With better synchronization between the release cycles of different projects, we can improve consistency of the user experience, more rapidly deprecate duplicate functionality, and increase inter-library compatibility.

Strategic Plan

The strategic plan will identify core needs and future challenges of the scientific Python community. For instance, rather than describe technical challenges that a specific project such as NumPy will tackle in the next decade, the strategic development plan would discuss general challenges that the array (or tensor) object will face (e.g., different architectures such as GPUs or TPUs). Individual projects could then decide which of these challenges to address, and decide on appropriate tactics. The plan will also be used by the community for support when applying for federal grants.

Our community has never before engaged in a deliberative process to try and anticipate future challenges. For the strategic plan to reflect the diverse interests and needs of the increasingly large developer and user community will require a multi-year effort with intense community engagement and dialogue. Much of this effort will first take place within individual communities and projects. Community concerns will then be collected and organized at the annual meeting, which provides a framework with which project leads can outline community discussions. We will also host events to engage with the larger community directly; for example, we will have

townhall meetings at annual events such as the SciPy conference.

Grant Proposals

Help projects prepare for future funding opportunities. This includes setting up the project for fiscal sponsorship (putting in place all required documentation and agreements, and clearing technical debt), identifying potential funding sources, and helping the projects to structure and prepare their grant applications.

***Activity:** Advisory Council Engagement*

To guide the direction of this work, help set priorities, and advocate for project adoption of our work, we will form an advisory council: a small group of experienced members of the community, who have founded and grown their own projects.

Council members will be invited to visit BIDS several days per year for us to get advice on shared engineering infrastructure challenges, as well as understanding their project-specific needs. We will also engage with these project leads throughout the year via regular conference calls, email exchanges, and project development infrastructure such as GitHub.

The council will help us select and recruit the 50 community leaders to attend the annual forum.

***Activity:** Annual Forum*

The main purpose of the annual three-day forum will be to guide the creation of the strategic plan and to help the community prepare grant proposals.

YEAR 1 We will send participants a detailed description of what we wish to accomplish with the strategic plan as well as a number of questions for them to consider before the meeting. We will ask community members to prepare talks outlining their individual views and visions. After general discussions, we would agree on basic tasks for individual software projects to conduct over the next year.

YEAR 2 Before the second meeting, we would draft the strategic plan based on the feedback from the first annual forum, our site visits with major science and industry user communities, our meetings with the advisory council, our interactions with scientists and students at Berkeley, and continued engagement with the developer

community. The second year would involve talks, developing shared grant proposal ideas, and group writing.

YEAR 3 By the third year we would finalize the strategic plan for the community for the next decade of development and growth, and continue working on new grant proposals together.

Activity: *Site Visits*

We will also engage with major science projects with site visits. For example, we will visit science teams involved with LIGO, LSST, STScI, and based at national laboratories such as Lawrence Berkeley and Los Alamos. We will also connect with industrial users in Silicon Valley, such as the Google teams working on Jax and TensorFlow. During these visits, we would work to understand the needs and challenges faced by these user communities.

Notable efforts in this area include:

- NumFOCUS: This 501(c)(3) was established to provide financial and legal representation for scientific open source projects. NumFOCUS partially addresses all of the issues above. This project differs from that of NumFOCUS in that it is specifically focused on planning for scientific Python (the annual NumFOCUS summit is much broader in scope). It is complementary to NumFOCUS's effort in that it makes projects aware of this support mechanism.
- Quansight: Quansight aims to connect industry partners with the open source software that they use. Quansight also aims to support open source development directly through Quansight Labs, and to provide branding support (both for projects and companies that use/support open source).
- Grants by the Moore and Sloan Foundations supported full-time developers on the NumPy projects, allowing that project to recover from years of technical debt, and to plan, along with the community, new and upcoming features.

Outcomes Table

1. **Projects share improved common infrastructure & process**
   (a) Re-usable libraries with commonly used functionality
   (b) Shared build system for binary binary packaging
   (c) Integrated websites and expanded documentation
   (d) Shared benchmarking
   (e) Developer Operations (DevOps)
   (f) Python 3 modernization
2. **Core projects are well coordinated**
   (a) Community best operating and development practices manual
   (b) Coordinated release schedule
   (c) Regular cross-project communication
   (d) Joint governance structures
3. **Community has a shared development vision**
   (a) Videos archive of project challenges, strategies, and visions
   (b) Community vetted strategic plan
4. **Community receives more funding to more individuals**
   (a) Grant proposals from multiple PIs
   (b) Community list of grant opportunities
   (c) Record of submitted proposals & archive of funded proposals