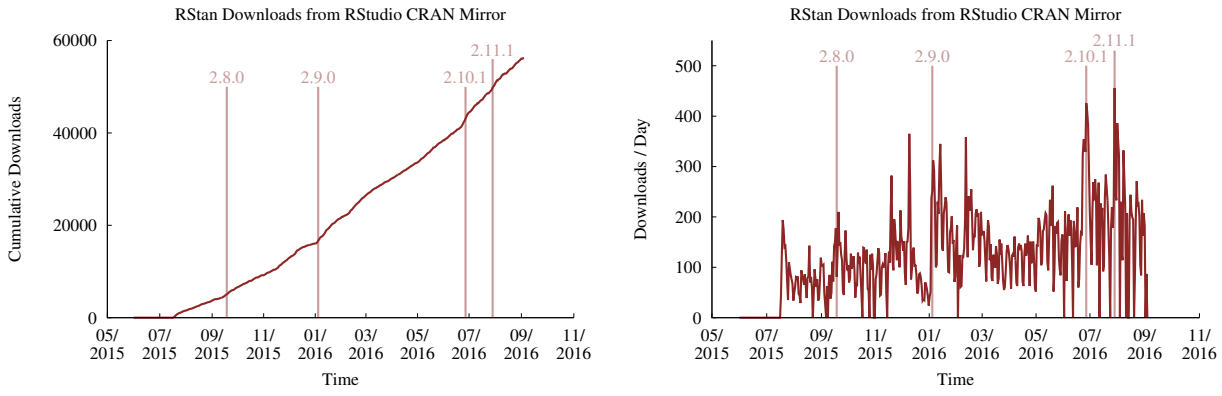
Stan

Fueled by technological and methodological advances, scientists, engineers, and business analysts are collecting more and more complicated data. To learn from that data, however, we have to build statistical models of the experiments from which the measurements were made and then then identify the configurations of those models that are consistent with the measurements. In particular, to accurately quantify our uncertainty we need to identify not just some consistent model configurations but *all* consistent model configurations. Fortunately, recent advances in statistical computing have revolutionized our ability to build robust statistical analyses in these complex problems.

Stan is a statistical library that facilitates general modeling, analysis, and prediction. Users first specify their models with a probabilistic programming language from which they can generate inferences using

- full Bayesian inference with scalable Hamiltonian Monte Carlo
- approximate Bayesian inference with automatic variational inference
- penalized maximum likelihood estimation with optimization.

These inferences are built on top of a powerful C++ math library that provides differentiable probability functions and linear algebra. Additional R packages provide expression-based linear modeling, posterior visualization, and model validation.



(a) (b)

Figure 1: Exact download statistics are limited, but we can gauge the overall growth of the Stan community by looking at subgroups that do offer statistics, such as users who employ Stan through RStudio. The (a) cumulative and (b) differential downloads indicate sustained growth of this subcommunity and, hence, the entire Stan community.

Stan has already had a very broad impact; thousands of people are using Stan to fit models, and scores of scientific papers have been written using Stan; a Google Scholar search for Stan's home page shows citations in clinical drug trials, general computational statistics, entomology, opthalmology, neurology, sociology and population dynamics, genomics, agriculture, psycholinguistics, molecular biology, population dynamics, materials engineering, botany, astrophysics, oceanography, election prediction, fisheries, cancer biology, public health and epidemiology, population ecology, collaborative filtering for recommender systems, climatology, educational testing, and natural language processing.

Because Stan is distributed through multiple channels that do not share detailed statistics, estimating the total number of active users is subtle. We can gauge the growth of our community, however, by studying a subgroups such as those who download Stan through RStudio (Figure 1). The sustained growth of this subcommunity indicates that the entire community is prospering.

Stan is freedom-respecting, open-source software (new BSD core, some interfaces GPLv3) that is associated with NumFOCUS, a 501(c)(3) nonprofit supporting open code and reproducible science. All of Stan products have been released under the most generous open-source licenses possible. Stan?s licensing satisifies the goals listed in the solicitation for software sharing:

1. it is freely available to everyone, not just researchers;
2. it is licensed so that it may be freely extended, customized, and incorporated into the maximal number of other tools;
3. the open-source licensing allows continuation of the project in the event of the original developers not being willing or able to;
4. researchers are free to modify the official versions of the software released by the Stan development team;

5. integration of user-provided code back into the core product for bug-fixes, examples, and enhancements is carried out using GitHub pull requests with integration testing.

Testimonials

Prof. Joseph A. Formaggio
Physics Department, Massachusetts Institute of Technology

Stan has been an indispensable tool in the analysis of complex data by our group at MIT. I insist that all our incoming graduate students learn to use it when they join our group, knowing that it can be quite powerful in whatever analysis they will later undertake in their time here. Our most recent paper, "Violation of the Leggett-Garg Inequality in Neutrino Oscillations" (Phys.Rev.Lett. 117 (2016), 050402) greatly benefited from having Stan as our analysis tool.

Nathan Sanders
Senior Director of Quantitative Analytics, Legendary Entertainment
(Formerly Graduate Researcher, Harvard-Smithsonian Center for Astrophysics)

I first encountered Stan while completing my PhD in astrophysics right after it hit version 1.0. It immediately had an impact on my research. Because the Stan modeling language is so intuitive and the NUTS sampler is so robust, it became the first modeling tool I reached for. It allowed me to easily transition classical models in my field from a maximum-likelihood to a full Bayesian framework, and to innovate new and complex models while focusing on scientific inferences rather than sampling algorithms and execution efficiency. My collaborators and I built a hierarchical model for supernova light curves, thousands of data points representing the brightness of exploding red supergiants at different wavelengths as the explosions evolve over time, that enabled us to directly model the population parameters of the progenitor stars of these explosions while accounting for a variety of observational biases like censoring, truncation, and selection effects.

Today, I use Stan as a primary mechanism for my work in industry. It's a terrific tool for our data science team here because the Stan language makes it straightforward to share human-readable models that can be executed across environments with minimal dependencies, and then adapted or improved with ease. The ecosystem of tools emerging around Stan, like rstanarm, shinystan, and loo, continue to make our work both more fluid and more effective over time.

Ellie Sherrard-Smith
Research Associate, Imperial College London

Stan has allowed us to fit a probabilistic model to data from a complex direct feeding assay experiment incorporating multiple mosquito per mouse biting rates and transmission cycles. This has provided a method to assess the effect size and efficacy of transmission blocking vaccines and pre-erythrocytic vaccines that can be used to eliminate malaria. The method provides a statistical means to capture the uncertainty in efficacy estimates for different combinations of treatments simultaneously (Sherrard-Smith et al. in prep). The range of statistical distributions available in Stan, such as the zero-inflated negative binomial distribution used here, enables a more precise population estimate of parasite intensity. Many of the processes governing parasite transmission are density-dependent (Shaw & Dobson 1995,

Shaw et al. 1998) indicating that there is huge scope for Stan to be a valuable tool for parasite ecology. Stan has also been used to fit non-linear logistic functions to binomially distributed data to assess the impact of insecticide resistance on public health (Sherrard-Smith et al. in prep). The versatility of the tool means that it has far-reaching applications with potential to transform our capacity to interpret data.