Budget Justification Subcontract Columbia

Fringe rate 27.1%, Indirect cost rate 60%

Benjamin Goodrich

Benjamin Goodrich, PhD, (6 calendar months) Co-principal Investigator, Lecturer in Discipline, teaches in the Political Science Department, the Quantitative Methods in the Social Sciences master's program, and the Environmental Science master's program at Columbia University. He will contribute 50% of his salaried time to the project through a subcontract.

With a Doctor of Philosophy in Government and Social Policy from Harvard University, his postdoctoral research experience in applied statistics and ample experience in developing and maintaining software packages, Dr. Goodrich has the required expertise and experience to lead with project with a special emphasis on the statistical and software/computational implementation aspects and issues of the research project.

Benjamin Goodrich has closely collaborated with Dr. Gelman, co-investigator in this proposal and his research team in the development of *Stan* and *Rstan*. *Stan* and *Rstan* are the probabilistic programming language and its interface in the R statistical programming environment, respectively, which the software product, this project proposes to develop is building on and developing further. Drs. Goodrich and Andreae closely collaborated with Jonah Gabry to produce the prototypes *rstanarm* and *shyinstan*, which will be further developed and hardened during the award period. Dr. Goodrich participated in other NIH funded research projects as a co-investigator.

Jointly with Dr. Andreae as co-principal investigator, Dr. Goodrich will have overall responsibility for all aspects of the project - scientific, administrative and financial. Drs. Andreae and Goodrich will prioritize research and development questions, analyze published and unpublished reports, write papers, meet frequently with the other investigators and supervise the administrative and data personnel.

As co-principal investigator (together with Dr. Andreae), Dr. Goodrich will provide oversight of the entire project, development and implementation of all policies, procedures and processes. In these roles, both will be responsible for the implementation of the Scientific Agenda, the Leadership Plan and the specific aims and ensure that systems are in place to guarantee institutional compliance with US laws, DHHS and NIH policies including human subject research, data and facilities.

As lecturer in the Political Science Department, the Quantitative Methods in the Social Sciences master's program, and the Environmental Science master's program at Columbia University, with ample experience in giving *Stan* and *Rstan* related workshops, he will lead the seminars and workshops to disseminate the software *rstanarm* and *shinystan*.

Andrew Gelman

Dr. Andrew Gelman is Professor of Statistics and Political Science and Director of the Applied Statistics Center at Columbia University. He will serve as a co-investigator on this project and will contribute (0.6 month) 5% of his salaried time to the project through this subcontract.

Dr. Gelman earned his PhD in statistics from Harvard University in 1990 under the supervision of Donald Rubin and has become a prominent expert in Bayesian hierarchical modeling, having won among others the ASA Outstanding Statistical Application Award. Dr. Gelman's team developed *Stan*, the probabilistic programming language, this project is proposing to develop further with *rstanarm* and *shinystan.* He supervised Jonah Gabry and Michael Andreae, when they initially conceived and build the precursor software of *shinystan.* Dr. Gelman and his research team closely collaborated with Dr. Benjamin Goodrich, (principal-investigator in this proposal) in the development of *Stan* and *Rstan*. Dr. Gelman is the author of standard books on Bayesian hierarchical modeling and on teaching statistics.

He will contribute his expertise in hierarchical modeling to the development of *rstanarm*. In particular, he will advise on parsimonious programming of the *Stan* model library called by *rstanarm*, but written in the programming language Stan and on the choice of default priors. A renown and sought after speaker at national conferences, he will participate in the dissemination of our products in workshops and seminars at national conferences. With his remarkable record of interdisciplinary collaboration and his prominent role as the founder of the graduate program at Columbia teaching Quantitative Methods in Social Sciences, his will contribute his expertise in teaching.

Michael Betancourt

Michael Betancourt, PhD, Consultant ($50,000) is currently a Postdoctoral Research Associate in the Department of Statistics at the University of Warwick, UK. With a PhD in Physics from the Massachusetts Institute of Technology, his research bridges theoretical and applied statistics, including the development of novel computational algorithms and robust statistical tools for data scientists. Since 2012, he has been a core member of the team around Dr. Gelman, co-investigator on this project, and instrumental in the development of *Stan*, the probabilistic programming language and its underlying computational algorithms. Our project to develop *rstanarm* and *shinystan* is building on this existing software *Stan*.

With his particular approach of understanding statistical model through their geometry, Michael Betancourt will be contributing to the development of novel approaches to diagnose non-convergence of Hamiltonian Monte Carlo simulations. By improving the visualizing of shape and correlation of model parameters and the resulting loglikelihood in *shinystan*, he will contribute to simple troubleshooting algorithms to assist data scientists in the optimization of performance and model implementation. He will also contribute to automated tuning heuristics and optimized model formulation of the Stan models library made accessible to data scientists through *rstanarm*. He will participate in the programming of optimized *Stan* models, called by *rstanarm*. Michael Betancourt will also help devise novel visualization tools and graphical interfaces in *shinystan*, our software package for interactive graphical exploration of *rstanarm* model output and MCMC simulations.

Michael Betancourt has ample broad experience in teaching *Stan* to a wide variety of users. This and his intense multidisciplinary interactions with biomedical data scientists ranging from malaria epidemiology to pharmacokinetics will be key in promoting advanced modeling in data driven outcomes research through principled and consistent, but accessible documentation, well maintained mailing lists, life meetups, and interactive workshops.

Jonah Gabry

Jonah Gabry M.A., Statistician at Columbia University and Stan core developer, earned his master in quantitative methods in social sciences at Columbia, where he started his collaborations with Drs. Andreae, Goodrich and Gelman on *shinystan* and *rstanarm*. He is a core member of the *Stan* team around Dr. Gelman, co-investigator, the software this project proposed to develop further with *rstanarm* and *shinystan*. He programmed most of the underlying algorithms, functions and routines of *rstanarm* and process underlying as well as the interface of *shinystan*. Jonah Gabry will be contributing 50% effort (6 months) of his salaried time to the project.

Jonah Gabry will be contributing to the development of the existing and the writing of novel functions, routines and algorithms for *rstanarm*. He will implement new graphical tools and interfaces, novel analytical tools for *shinystan* in close collaboration with Michael Betancourt.

Travel

Travel funds are needed for local and intercity travel to conference sites. This includes funding to send 8 investigators to teach about two software workshops per year and present project results at national meetings, as justified below and distributed across the budgets of both collaborating institutions.

Outreach and dissemination are major components of the proposed project to build accessible software packages for clinical data scientists and promote their widespread use. One part of the dissemination plan is to conduct free hands on workshops at national conferences with data scientists, who are likely to further propagate the use of our new software.

The two workshops per year will be hands on practical exercises to introduce prominent data scientists to the new software we developed. We will carefully select and invite ‘multipliers’ to attend our workshops for free; by ‘multipliers’ we mean academics, teachers and prominent researchers in the clinical data science community who themselves will train other users in our software or propagate the software use on blogs or in publications.

Four instructors will attend for each of the two workshops per year, as the workshops will be conducted in small groups with a student to teacher ratio of 1:5 and will have with about 20 participants each for maximal interaction and participant engagement. Workshops will be offered at the Joint Statistical Meetings of the American Statistical Society, the Statistical Association Conference on Statistical Practice, the Conference and Workshop on Neural Information Processing Systems (NIPS), the Annual Meeting of the American Society of Anesthesiologists, the Annual Congress of the Society of Critical Care Medicine and the American Thoracic Society International Conference, among others…

Online Software Repository Github

Online software repository: $2000 per year are budgeted for the Github online repository fee to collaborate on a private repositories during the initial design and development of the software and its components. Eventually the software will enjoy free public access also through the Github repository and finally will be published as a package on CRAN the public R software and package repository.