Our aim is to radically lower the barrier of entry to high-quality modern teaching of data science. Our team and resources each play their role in removing current obstacles.

These obstacles are clear by example. Berkeley currently runs a highly successful data science education course with many core features of the courses we propose here. This is no coincidence; we have been strongly influenced by their model. But - they built their course with collaboration of multiple teachers from their statistics and computer science department, they have significant funding and support from the university, and they are building on top of tools like JupyterHub, where many of the main authors are based in the university. To run their platforms, they have a dedicated very high quality software development / systems operation (devops) team. They already have strong links with EdX, the company and platform. New teachers writing materials have the support of the development team there, to write new notebook with testing and other pedagogical scaffolding. Unfortunately we cannot use their teaching materials, because of their restrictive license.

All of these are currently very difficult to replicate, even for universities with significant resources, let alone individual teachers who have discovered the transformative power of this way of teaching. Our costs are all aimed at removing these barriers, so it should become practical for individual teachers to adopt, adapt, and deploy very high quality data science teaching materials, in various course forms.

To start with devops costings; our consultancies are with two key players in making the standard teaching technology more accessible. 2i2c.org are a company founded by many of the key players who have built the Berkeley infrastructure for running their notebook services so they can scale to many thousands of students. Through them we have access to the best and most experienced devops expertise available. They have a specific remit of making this hosting more accessible, and easier to deploy. We will be working with them on integration with Open edX, the virtual learning environment - work they have already begun with Berkeley's own data science edX course. We will also work with them to make it easier for others to use our preconfigured JupyterHub hosting, so others can start using this technology more quickly and more cheaply. Finally, we will will working with them on making it easier to deploy standard instances of this hosting technology with no or very little support from us, or from technical staff.

Similarly, our contract with Sasu Sasu Nuli is with the primary open source developer for automated, configurable deployments of the standard Open edX VLE. This developer has built one of the standard deployment mechanisms, that allows one-click or similar deployment to cloud services, using the standard Kubernetes cloud management system. His work is in the process of being discussed within the Open edX community for adoption as the default installation mechanism. He has enormous experience developing within Open edX, so he is the ideal choice for collaboration with the JupyterHub team, above. These two companies give us the best possible chance of having simple or even one-click installers for full integrated Open edX / JupyterHub systems that will scale to thousands of students.

In order to test this work in our own courses, and show it in action to others, we need to be able run hosted instances of these systems. We therefore ask for hosting costs in terms of virtual cloud machines, and for consulting for hosting support, to maintain the hosting systems with high-uptime, and on-call support.

The work on technology will be much less valuable if we do not have materials for essential data science courses. We have argued that the core of data science for research is **reproducible data science** - a *code-to-learn* basis, on which we build *code-for-reproducibility* - but we also want to show that others can build on these foundations to teach specific

application areas and domains, such as probability and statistics.

Because the current requirement for expertise is so great, we need senior, experienced teachers to build and review the material, and who also have very substantial experience of using code and contributing to open source. The PI, Matthew Brett, is one of very few teacher / developers in the life sciences who is an experienced researcher as well as being a very experienced developer, with a long history of contribution to open-source, and leading open source projects. He is one of very few teachers who have taught a semester course on reproducible data science. We have published a description this course; it was very successful in engaging students in reproducible research practice [1]. Because of his deep experience of Python programming, he has many contacts in the open source community, and has already done work to make it easier to automate marking and build exercises, all available on Github.

Juan Klopper has demonstrated quite extraordinary commitment, deep knowledge, and long, fruitful experience of remote learning, as he has led in teaching and research in the biosciences, while being a very busy full-time surgeon. He is an award winning teacher with three major courses on Coursera, and he has led his university - and continent, in online learning in the health sciences. Given the truly remarkable amount of teaching he has done, with so little time to work on it, we fully expect to get extraordinary value from buying 50% of his time.

The other established faculty on this grant each brings very great experience of various aspects of data science and statistics teaching, and development of remote learning. Peter Hansen has an unusual depth of experience in computing and academia; he knows a wide range of tools and languages, and runs a data-intensive teaching module that fits well with Brett's prior experience, so he is an excellent partner for work on the core courses, but particularly the reproducible science course, with neuroimaging as the initial example application.

Federico Turkheimer has very long experience of teaching statistics to undergraduates and post-graduate students, with a neuroscience background, including writing his own textbook. Brett and Turkheimer have worked together very fruitfully in the past, and we expect this to continue for the work on this grant. Brenda Williams has very deep experience of running distance learning courses, and has been highly successful in building the King's distance learning programme. She is well-suited to the role of academic lead taking responsibility for our use of the VLE and student interaction. Robert Leech combines much experience of using code, and teaching with code, and can combine a strong research and teaching background with the ability to contribute to the machinery of the VLE, code contribution, and other aspects of improving process.

We have also asked for funding for a full-time post-graduate teaching fellow. We think this is essential to maintain the health of this community and help to expand it beyond the current funded investigators. Our skills are currently very rare, and hard won. There are few, if any, training courses to acquire these combination of skills. We will mentor the post holder, with each senior fellow taking some time train them in their respective expertise. We believe this breadth of training will be unique, and we expect the post-holder to make important contributions to the current transformation in data science teaching, after the grant has finished.

1. Millman KJ, Brett M, Barnowski R, Poline J-B (2018) Teaching computational reproducibility for neuroimaging. Frontiers in Neuroscience