Data Driven Discovery Investigator program grant proposal

I've just made my full application to the Moore Foundation's Data Driven Discovery Investigator program available, but wanted to post an HTML version, too. You can also see a short sci-fi story about what I want to enable.

Tue 13 May 2014
By C. Titus Brown
In science.
tags: research out on a limb moore

You might wonder why I'm posting this. Well, there's a snowflakes chance in heck of this sort of thing being funded federally, so even if I *weren't* in the habit of posting my own grants publicly, I would probably post this one; I'm not going to be submitting it anywhere else.

Moreover, there are ideas in here that I'd rather not take to my grave. (Bill Tozier gave a really sobering lab meeting on how many scientists leave behind thousands of pages of "good ideas." Not sure these are good, but *I* think they are.) If they inspire someone else to do something good, then so much the better. Just... make it open source, m'kay?

Instructions

Please submit a three-page supplement to address the two questions below:

1. What do you envision as the five-year impact of your work on one or more of the natural sciences? Is there a key, fundamental question that you are trying to answer? How will you measure progress towards answering this question over the five years?
2. How will you advance data science methodologies, such as statistics, machine learning, automated inference, etc., to achieve this goal? We are particularly interested in the ways that the data science methodologies that you propose to develop can be applied to other fields beyond the one you focus on and shared. Please discuss these plans. What work products do you plan to make open source?

In response,

Summary of proposed work: software to support biological inquiry

Biology is increasingly making use of large scale sequencing of non-model organisms to inform ecological, evolutionary, and developmental research. However, we lack the basic infrastructure to collaboratively store, index, search, and mine these sequence data -- each lab's data is generally an island unto itself, and often cannot be queried even within the lab.

I propose to build a lightweight graph-query system for gluing databases together in a distributed manner (see Figure 1). This project, built on the pygr sequence graph database system, would enable labs to work with their own sequence data, make it available, and search and link across databases and data sets in a lightweight, federated way. We would build turnkey open-source software implementing the database backend, with easy virtual machine/cloud setup instructions and tutorials, such that labs could quickly and easily create their own sites. This software would consume the output of existing cloud-enabled analysis pipelines, including output from Galaxy, IMG and IMG/M, MG-RAST, and our own khmer-protocols.

While the core ideas already exist, we would connect them together to support our own and others' scientific inquiry. In particular, we want to support *automated* data exploration in ways that are simply not possible today; of particular importance, this would enable more sophisticated data mining approaches than the field currently uses.

Openness: Everything we do is open source, open access, developed on a public versioning site such as github, blogged about, and discussed on social media. All of our papers will be highly reproducible, with completely automated build scripts and figure-generating notebooks placed online with no access or reuse restrictions.

Five-year impact

Our five-year impact would be built on three deliverables:

1. a distributed graph database system for improved data analysis for core biological research.
2. public sets of homology and functional predictions based on public metagenome and transcriptome data, stored in a graph database server.
3. an open, scalable, and automated system for building the prediction databases, using cloud servers.

The impact itself would come from two different components of the project, both built with an eye to sustainability: first, the organization and public availability of useful data with an interaction interface; and second, the demonstration that such a distributed system provides useful, sustainable functionality. As part of this we would hope to see an active and growing community involved in further development of the technology, as well as many users using it.

Fundamental scientific questions

My lab works on data-intensive biology, where our emphasis is on using large DNA sequencing data sets to generate and refine hypotheses. We work on algorithms and software that enable experimental biologists to tackle a variety of fundamental problems.

The specific challenge identified in my preproposal is the challenge of assigning function to unknown genes in both metagenomes and transcriptomes. However, my interest is broader: I plan to build a framework that would allow progress on many different problems in biology, driven (at the start) by those problems that my lab is currently working on:

- *Function in microbial "dark matter". Assigning putative function to sequence from complex microbial communities is incredibly challenging. In support of building putative functional assignments, we need a rich query interface that lets us search for correlations between gene presence and metadata characteristics across databases. We are already working with soil, sediment, symbiosis, marine virome, and hot spring data.*
- *Function and phylogeny of genes in eukaryotes. As with shotgun metagenomics, transcriptome sequencing has become commonplace, but we need tools to support downstream inquiry. We have few tools to link homology and gene structure across many transcriptome sequencing data sets, and no standing databases that I know of. We are already working with animal transcriptome and marine eukaryote data to answer questions about vertebrate gene evolution, metabolic gene function, and ascidian and cephalopod evo-devo.*
- *Linkages between microbes and the big city environment. In September, I will be joining NYU CUSP for a year to work on the data integration aspect of the NYC Metagenome Project. We will be tracking microbial biogeography through the city and inferring microbial function across sewage, money, subway cars, and air.*

These are all *data integration* problems, where we fundamentally lack not only the algorithms and approaches but the perspective to tackle them effectively. Because of this lack, there is a growing amount of data locked up behind lab walls, awaiting publication; I believe that by providing effective and functional database publishing and query approaches, we can help unlock this data.

Measures of progress: 5 years out

Progress towards answering these scientific questions, and building sustainable infrastructure to help us and others continue to do so, can be measured in two ways. The first measure is traditional: publications. In five years I plan to have several publications using our software to integratively analyze data across environments and samples. These peer-reviewed publications would demonstrate the effectiveness of our tools and approaches in the only way that many scientists will respect. These publications will be placed on preprint servers for pre-pub peer review, made open access, be highly reproducible, and will openly provide data and source code.

The second measure of progress is less traditional but perhaps more important: if we are providing important and useful solutions that help address important scientific questions, our techniques should be be adopted by others. Already our data structures and algorithms have served as a foundation for new approaches; the diginorm algorithm has been incorporated into several widely used assemblers; and our khmer software is quite popular, with thousands of downloads a month. The recent release of khmer v1.0 has also seen a substantial increase in community participation in our software development. This adoption of our software has been driven by multiple factors, including effectiveness of the approaches, engagement with the community, and quality software engineering approaches. For this proposal, I would hope to see similar adoption of our core graph and server technology within 5 years, with dozens of labs running their own servers and making their data publicly available.

Advancing data science methodologies

Although the pygr graph database software itself is sequence focused and unlikely to be adopted outside of biology, the approach and perspective are broadly valuable.

Perspective shift: planning for poverty: Most current cyberinfrastructure development efforts rely on substantial sustained funding to a centralized authority. This renders them vulnerable to funding lapses, budget cuts, and leadership transitions. More decentralized and open bottom-up cyberinfrastructure models have not gained much traction in science.

If this project succeeds, it will succeed in large part *because* my lab uses bottom-up approaches: we will *start* with open source, open development, cloud computing, open community interaction and participation, and training in support of our scientific and software goals. This also increase our "bus factor" and makes our work much less vulnerable to funding lapses or loss of project leadership. This model of "planning for poverty" has been explored in science -- but largely due to failure to attain grant renewals. Here we are using it as a planned strategy that can be elaborated through experience.

Perspective shift: investigation of federated infrastructure: As with development, centralization of infrastructure and process is a central point of failure. Most database and Web site efforts lapse with funding, delaying scientific progress. Moreover, this emphasis on centralization means that database hosting often relies on "big iron" resources that are not widely available.

Our proposed graph database overlay does not rely on central data servers, which are a serious failure point in the era of Big Data. While the pygr project currently relies relies on a centralized namespace, this could be refactored to use a decentralized authority scheme (e.g. blockchains). We plan to enable any lab to quickly and easily make their data available in the cloud for linking and query, and provide push-button migration mechanisms to push data to archival locations such as figshare. This would make it technically easy to share data in a decentralized way.

This project also rests on our existing efforts to build open (and tested) computational data analysis protocols on cloud infrastructure, thus building sustainable process on top of publicly available infrastructure. Our cloud-enabled protocols for metagenome and transcriptome assembly and annotation can be run on expensive "medium iron" cloud resources. We plan for limited resources, provide explicit execution instructions, and test our materials regularly.

Federated infrastructure provides a sustainable path to the future. We hope to demonstrate that one can be successful in science with this model.

Building better computational scientists through training:

As part of my project, I would explicitly support computational training activities in my lab. Training is already part of my lab's culture: more than half of my students are accredited Software Carpentry instructors, and, in large part due to my encouragement, MSU has more accredited instructors than any other institution in the world. We regularly run Software Carpentry and other training events (5-10/yr) across all levels of expertise and across multiple domains. These workshops emphasize critical thinking about computational science, teach version control and testing, and introduce students to methods that encourage reproducibility. As part of the larger Python and R scientific training communities (that several other second-round DDD applicants also participate in), we foster better practice in this generation of scientists and help train the next generation of scientists, with obvious opportunities for broader impact on data science across all fields.

Concluding thoughts.

Data intensive biology is in need of different tools, perspectives, and infrastructure that support queries across distributed data sets. It's long past time to start building these tools and trying out new perspectives.

Post submission re-read thoughts:

1. Ok, look, you try to write something comprehensive in 3 pages :). This is distilled from 5-10 pages of notes, text, and discussions.
2. Can you spot the typo? Hint, it should be **in** expensive. Oops?
3. The competition on this one is going to be brutal. We've been told to expect a relatively high 15% funding rate (better than many NSF and NIH proposals) but I know several of the other second-round applicants and they're all *really* good researchers. I'm also submitting into an area that Moore has a lot of bitter experience with, and so I expect my proposal will be reviewed by people who know where all the likely failure points are. Yay?
4. If you have skeptical comments or questions or thoughts, please comment or tweet! I'd rather hear 'em from you first than from the Moore Foundation people.
5. Reminder to self: look up bio4j. :)

--titus

Comments !

(Please check out the comments policy before commenting.)