A conversation with John Wilbanks on 03/25/13

Participants

- John Wilbanks — Chief Commons Officer, Sage Bionetworks
- Alexander Berger — GiveWell, Senior Research Analyst

Note: This set of notes was compiled by GiveWell and gives an overview of the major points made by John Wilbanks.

Summary

John Wilbanks is the Chief Commons Officer at Sage Bionetworks and previously worked at Science Commons. GiveWell spoke with him as a part of our investigation of opportunities to improve biomedical research. The main subjects that were discussed were Sage Bionetworks and issues surrounding open data in science.

Sage Bionetworks

Sage Bionetworks is a research institute in Seattle that promotes biotechnology by practicing and encouraging open science. Its funders include the state of Washington, through a fund from the tobacco settlement, and the National Cancer Institute. It has a staff of about 35 people.

Sage Bionetworks was spun out of Merck, the pharmaceutical company, as a nonprofit. Their core technologies are machine-learning algorithms for using genetic information to predict health outcomes.

Sage Bionetworks' facilitation of data donation

John Wilbanks has served as the chief policy officer at Sage Bionetworks, designing systems that let people donate their data to be used for computational research. There is not currently an effective system to allow individuals to donate their data to science. There are significant difficulties with the privacy, updating, and sharing of individual medical data. Sage Bionetworks is working to create a repository of data, which will be updated over time, for researchers to use for a variety of purposes. This kind of database has been very successful in other areas and doesn't yet exist in the medical space.

Open data in the life sciences

The lack of professional incentives for scientists to share their data

Although scientists already collect a lot of high-quality biomedical data, they have no

incentives to share it:

- By sharing data, they may forego opportunities to publish papers based on the data in the future.
- Scientists don't receive much or any credit for discoveries that others publish on the basis of data that they shared.
- It is often time-consuming and difficult to adequately document and prepare data to be shared, and scientists aren't trained to do so.

However, it doesn't take a very large percentage of a population to share in order to create a valuable resource. A very small portion of photographers use Creative Commons licenses and only 0.3% of people who use Wikipedia make any edit in year, but each of there proportionally small groups are large enough to make a big difference in their industries. If a small fraction of scientists were willing to share their data despite the lack of incentives, there could be disproportionate benefits for the entirety of science.

Funder efforts to improve incentives

The main candidate for a long-term solution to the problem of there being little data sharing is for funders to come together and require that scientists improve their data sharing practices. Working to convince big funders to adopt good open access policies is one of Wilbanks' top priorities.

U.S. government efforts to require more federally-funded research to be shared openly will have positive effects, partially because they will lead to the creation of an infrastructure for sharing, analogous to PubMed Central. Philanthropically funded projects will be able to piggyback on the federal efforts by incorporating the federal open access policies by reference and using the federal infrastructure.

Rather than funding specific research projects conducted by universities, philanthropists should fund the production of true public goods like high-resolution data on the populations of interest to them. The availability of such data would easily attract the best computational researchers at this point, but. in 5-10 years, there will be a huge amount of high quality data available, and it will be far harder to get the attention of computational researchers. On the other hand, many more PhDs will be data-savvy by then, and there will be many more data scientists in the marketplace.

It would also be better to fund the collection of more high-quality open data rather than narrow hypothesis-driven lab research. Most diseases that need further work are too complex to make progress on by trying out one hypothesis after another in the lab. It’s more promising to fund a project like following 10,000 people for a year to collect a large amount of medical data, which can then be mined, with all the tested hypotheses logged publicly and subject to later follow-up by other

researchers.

Virtually no funders are funding efforts to create this sort of large public dataset. The NIH doesn't have a program structure for funding the creation of this kind of science, which can be used repeatedly. It is not usually considered "good science" to do this kind of public good creation – you can't win a Nobel Prize for it.

Potential problems with data mining

The problem of "most observational epidemiology research findings being false" is unlikely to be solved by the collection of large public datasets, and may be exacerbated.

However, large public data sets are likely to be analyzed in a more statistically rigorous and transparent way, reducing the high level of false positives that characterizes current observational research. An effective "GitHub for data" might be particularly helpful with this, because it could show how the scientists reached the conclusions that they did, and might reveal instances where scientists over-fit their models to the data.

Infrastructure to support data sharing

There are number of aspects of infrastructure that are needed to cultivate an open data ecosystem:

- Version control for data and collaboration
- Sharing intermediate files
- Data citation (for both sourcing and giving credit)
- Compression and analytics

It's important that as these solutions are developed, the creators pay attention to user experience and design. Too often, non-profit projects are poorly designed and hard to use. They should also be open and interoperable.

Sage Bionetworks' Synapse software for version control of data

Sage Bionetworks' Synapse software is a solution to the problem of collaborative version control for the life sciences. It is meant to operate as a sort of GitHub for life sciences data, though it does not incorporate Github's citation/referencing system. It is meant to solve the narrower collaborative version control problem, which is often an issue internal to labs or companies, and is not necessarily meant to address the broader issue of attributing credit, which GitHub also does to some extent.

Usage so far has been driven by specific challenges that Sage has hosted, which have drawn hundreds of participants. One pharmaceutical company that Wilbanks is

aware of plans to install Synapse locally.

Science Commons

Wilbanks previously worked at Science Commons, which has since been collapsed into Creative Commons, which had previously owned it as a subsidiary. At the time, the organization was focused on two issues:

(1) Promoting open access scientific publications in general, and the use of Creative Commons licenses for scientific publications in particular.
(2) Finding research areas where the concept of Digital Commons wasn't being applied, but could be applied, and helping develop the technological and legal infrastructure to put the concept in place.

People for GiveWell to talk to

Kaitlin Thaney — The Manager of External Relationships at Digital Science, which is a subsidiary of the scientific publisher MacMillan Publishing. Digital Science serves as a venture capitalist for open science projects.

Dave Clifford — An independent consultant who has worked at PatientsLikeMe, a website for patients to post about their experiences with disease and connect with others who have the same disease, in the process generating data that helps entities in the health care sector develop more efficient products, services and care.

Lucky Gunasekara — The Executive Director of Vulcan Labs, and a member of the Pioneer Advisory Group at the Robert Wood Johnson Foundation.

All GiveWell conversations are available at http://www.givewell.org/conversations