Metcalfe's law and the biology information commons

Stephen H Friend & Thea C Norman

Open collaboration on biomedical discoveries requires a fundamental shift in the traditional roles and rewards for both investigators and participants in research.

"Following the light of the sun, we left the Old World." –*Christopher Columbus*

Maps foster discovery and chart human understanding. In the age of Columbus, accurate geographical maps made exploration, mercantilism and trade possible. In 2013, geographical maps are an essential collective resource and underpin our ability to model complex systems of climate ecology, migration, economics and sociology. When we consider the topography of human biology, what do our current biomedical maps tell us about our ability to discover and to understand? Why is it that modern biomedical research has not yet afforded accurate individualized maps of disease and wellness? What new course must we as biomedical researchers chart? And what shall we bring and what shall we leave behind so that we develop maps of health and disease that bring us to a new world?

Why current biological maps are antiquated

Biology is a complex system of thousands of interacting components and a very difficult system to efficiently map and model. Until now, answers to the questions "How do drugs work?" and "What do diseases do?" have proved elusive, mainly because of the lack of precise surveying and measuring tools. For example, clinical research data are most often collected at infrequent time points and from distinct groups and populations. Similarly, a doctor's diagnosis is based on abnormal threshold measurements, such as broken bones or hemoglobin A1C scores, and provided without knowing what scores constitute wellness for that person. Like one frame in a full-length movie of 200,000 frames, a single biological

New York Public Library / Photo Researchers, Inc

Maps will be just as important to biological discoveries as they were to the discoveries in the era of Columbus.

Stephen H. Friend is the President and a Co-founder of Sage Bionetworks and Thea C. Norman is the Director of Strategic Development at Sage Bionetworks, Seattle, Washington, USA.
e-mail: thea.norman@sagebase.org or friend@sagebase.org

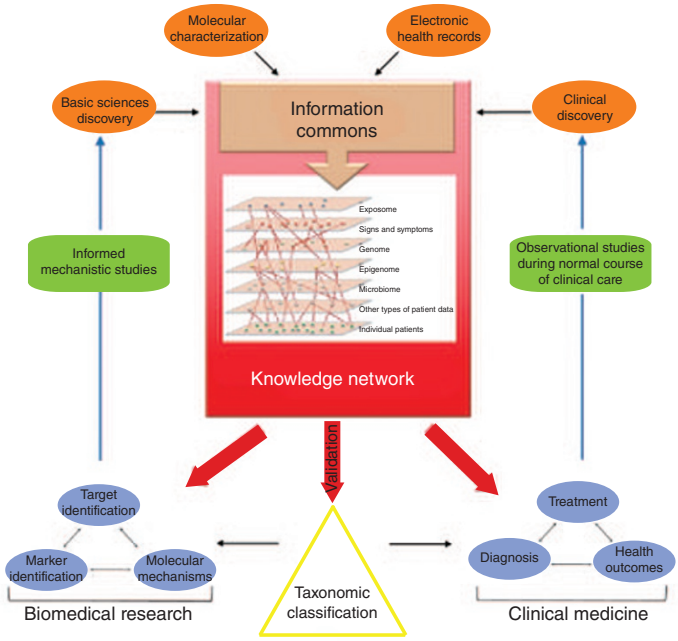

Figure 1 Depiction of an 'information commons' in which data on a large population become broadly available for research use. The information commons in the center of the depicted network contains patient data and current disease information. Patient data in the commons are directly linked to individual patients and are continuously updated as new data emerge from studies and routine health care. The continuous accumulation of data into the commons can be used to drive (i) disease classification; (ii) new clinical approaches (that is, diagnostic and treatment); and (iii) basic biological discovery. Source: Committee on a Framework for Developing a New Taxonomy of Disease.

reading is a frozen snapshot of a complex living system and a crippled approach to understanding the story of how biology works.

What's worse, drug developers often collect clinical data on carefully defined cohorts and populations and then use them to define how another completely different population (or the so-called average person) will respond to a new medicine. Indeed, our entire medical philosophy seems built on the erroneous assumption that averages can accurately reflect the benefit or risk of a drug, when in real life there is no such thing as an average patient. Unfortunately, as demonstrated by the frequency of recent post-approval safety issues, the law of averages often fails to predict the actual impact of prescription medicines on individuals.

And so, similar to the fabled blind men who touch an elephant to learn what it is like, today's ill-equipped doctors and researchers resort to using disease symptoms and pathology to define what it means to have a disease and to implicate targets and biochemical pathways. Not surprisingly, drug development statistics confirm what we might expect from such a blind approach. Every year, the US pharmaceutical industry increases its spending on research and development. In 2010, the outlay reached $60 billion. Yet the past 10 years have seen fewer major industry breakthroughs than we saw in the final decades of the 20th century (e.g., statins for heart disease, AZT for AIDS, and Herceptin (trastuzumab) and Gleevec (imatinib) for different cancers). Of the 37 new drugs approved by the US Food and Drug Administration in 2012, an increasing proportion (13) were for rare diseases, another 11 added only months of survival to patients with various forms of cancer and one drug (raxibacumab) has no patient population whatsoever (unless there is an anthrax outbreak). Consultancy Oliver Wyman has calculated that since 2005 the value generated by a dollar invested in pharmaceutical R&D has plunged more than 70%[1].

Of course, it's not just complex biology that contributes to the current 99% failure rate and high cost of new drug development. The legal and business decisions of pharmaceutical companies, archaic clinical trial design, more stringent regulatory climate and troubling global economic circumstances are all contributing factors. So, too, are the increasingly expensive technologies and assays that we leverage to characterize new disease targets and drug actions. Applied in the current paradigm, these new tools are not getting us closer to answering the question of whether modulating a given target in a particular biological context results in the amelioration of a disease phenotype in a human being.

Orienting research to embrace complexity

The past few decades have seen the emergence of genomic tools that for the first time allow us to measure biological systems precisely and frequently. Although it might have happened that our new tools confirmed our existing ideas about biology and diseases, it turns out that they are showing us otherwise. Our new tools are showing us that common diseases assumed to be relatively homogeneous, such as depression, breast cancer and even hypertension, have been shattered into a multitude of molecularly related but distinct disorders.

Not unlike the physicists' tools that in the late 19th century revealed a previously unknown and complex world of subatomic particles that eventually was rendered comprehensible by the Standard Model of particle physics, these new biomedical instruments have uncovered layers of complexity and ambiguity that in their depth of data have the power to eventually organize the current tsunami of new observations. Many assume that investing in 'omics' technologies, such as genome-wide association studies, and the cataloging of new genomes will, all on their own, be sufficient for us to make sense of the biological complexity we can now measure. But a genocentric approach is limited to a single data dimension and is unable to provide a complete enough context to see and understand a biological system in its entirety. For example, we have seen an epidemic in obesity and diabetes across the world in the past 50 years, a timeframe during which we can assume genetics has not changed substantially. Genomics by itself cannot tell us much about the incidence of obesity and diabetes. But when combined with other phenotypic data and risk factor information, such as a person's body mass index, diet, lifestyle and environment, DNA sequence variations that exert their effects in specific contexts can be identified[2].

We need to invest in the creation of causal knowledge network models and systems that correlate genomic data with a quantity and variety of individual in-depth phenotypic information that goes well beyond what is currently available in electronic medical records. We believe that the following four capabilities put us at a tipping point and historic moment of scientific opportunity to construct what the US National Academy of Sciences calls a "knowledge network" (**Fig. 1**)[3].

- The ability to leverage new internet and cloud computing technologies to take an open approach to biomedical problem solving.

- The ability to leverage open social media so that citizens, researchers and experts can interact in new ways to solve problems.
- The ability and interest of patients to control their medical data, to have access to affordable data collection tools and apps (e.g., FitBit and Moodpanda) and to have a voice in how the data get shared.
- The ability to integrate genotypic and phenotypic information to build a variety of top-down, bottom-up and middle-out models of disease.

With the above capabilities, and by aggregating and integrating stupendous quantities of genotypic and phenotypic information as a common resource, we believe that it is possible to replace today's symptom-based approach to health with a 'precision medicine' approach that is driven by data and focused on the individual.

Why we need to change

Our ability to shift from increasing complexity to distilling the patterns that determine health and illness is within reach and no longer limited by tools or technology. In light of this, we believe that the arc of our success to build a knowledge network will ultimately be determined by the way we work or do not work together.

For where we need to go, it is problematic that today's medical information system operates in a closed guild-like fashion where institutions, companies and researchers are too often rewarded for not sharing hinders progress. Siloed institutions flourish and compete for government funding, data are controlled by companies to protect intellectual property and by individual researchers to protect publication and grant-raising ability, and regulators in partnership with the pharmaceutical industry persist in treating clinical study reports as confidential documents.

As a result, scientific information that could help others is often sequestered for years. And even after publication, many commercial publishers place data and results behind their own firewalls that only paid subscribers can access. This approach of siloing information and waiting for a patent or paper before sharing is literally making us sick. We should be humbled by the complexity of the omics world we have just wandered into and ask ourselves if the emerging flood of observations requires us to fundamentally rethink how we practice medical research and the systems by which we share and disseminate the data.

Box 1 What is Sage Bionetworks?

The nonprofit Sage Bionetworks (Seattle) was started in 2009 out of a conviction that biomedical research will be more successful and affordable if it runs off of a layer of data and models in an open information commons that each of us and teams of teams can access to make better, faster and more relevant discoveries. Sage has been experimenting with governance and technological infrastructure that can support improved collaboration as well as rewards and incentives that motivate researchers to share and develop their ideas before publication, so that rapid learning can take place. Similar to how ARPANET enabled the internet, Sage's projects are fragile first steps toward a vibrant information commons that enables data-driven medicine to flourish. From spectacular failures to notable successes will come meaningful learning about which tools, incentives and approaches best enable open research and citizen involvement.

Sage's goal is to develop predictors relating to health and to accelerate biomedical research through open systems, incentives and standards. So that everyone can easily donate their data to an information commons where it can be worked on all the time, Sage Bionetworks, working with others, has been developing two key pieces of infrastructure: experimental consent tools and standards that allow citizens and patients to donate their data and control how those data are used, and a 'researcher's sandbox' called Synapse where the data are stored and disease models can be built.

Over the next decade, ever-expanding data will transform biomedical research approaches and accelerate healthcare discoveries through the use of predictive models that give insight into outcomes and responses to treatment. This advance will be best harnessed when individuals and groups can collaborate openly on discoveries, with a fundamental shift in the traditional roles and rewards for those involved. For example, there is a new openness on the part of funders to provide credit not only in terms of papers published but also based on the generation of data and data sets that many others use in their own work and analysis[4]. Additionally, discovery will happen faster and more efficiently when the existing practices and constraints inherent in research are removed so that laboratories can cooperate, rather than compete, and then benefit from the power of collaboration and from the reward of shared publications.

For this to be possible, we need to adopt the open practices of adjacent fields of science and engineering, such as astronomy, math, physics and software development. When those communities were faced with the question of how to approach the analysis of their own massive data sets, they learned how to share data and the resulting models as a common resource. Working from an information commons makes the complex modeling of climate, ecology, migration and economics possible. And one has only to consider the birth of the internet (the ultimate information commons) from thousands of interconnected open source software projects to realize that an information commons provides a wellspring of new ideas that make people more efficient, drive innovation and fuel successful private endeavors and academic research. Those communities have learned that a layer of shared information accelerates innovation and nurtures the development of commercializable private goods.

At Sage Bionetworks (**Box 1**), we are taking the first steps in building the type of infrastructure and communities that will be critical in creating a biological information commons. Achieving this aim will involve empowering patients to participate in research studies and provide samples; creating an open web-based compute platform to facilitate access to, and the sharing of, biomedical data; and assessing different pilot studies to both learn the drawbacks and benefits of our models and understand which approaches work and which do not. In the following sections, we discuss each of these aspects in turn.

Empowering patients in open research

In the current constrained funding environment and with today's wide availability of lab service exchanges, sensors, mobile phone apps and online tools, it is probably less expensive for citizens and patients to continuously collect and donate their data than it is to go on investing in private foundations and institutions to do this. And if this less expensive, high-dimensional phenotypic data collected by citizens and patients could be correlated to genomic information and aggregated in one place, then the same strong mathematics that drive big data analysis in physics, meteorology and financial markets can be applied to the analysis of which genetic variations in a population correlate to diseases, to drug response and to long life. The high dimensionality of the integrated data also provides a meaningful context for patient-reported data to be better understood and for noise and the so-called placebo effect to be identified.

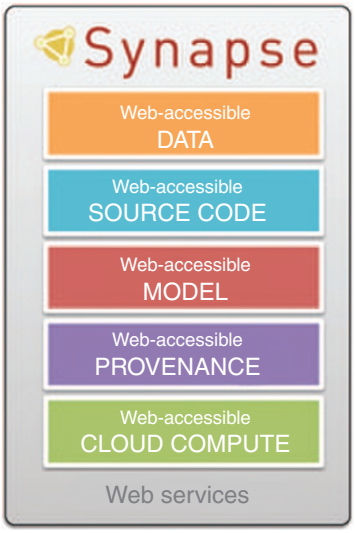

Figure 2 Attributes of Synapse, Sage Bionetworks' open compute platform. The Synapse platform comprises a set of shared web-accessible services that support a website designed to facilitate collaboration between scientific teams. On Synapse, scientists interact and share data, source code, models and analysis methods, both on specific research projects and across otherwise disparate projects. Provenance visualization tools allow users to formally express the relationships among resources in the system and better document and communicate about complex scientific workflows.

The problem is that right now, it's not easy to donate your data to an information commons, in part because current standards for patient consent are not geared to open research. Most often, patients sign a consent form at the beginning of a clinical trial; these forms usually limit what data will be generated, how they will be used and who can look at them. Although current Health Insurance and Portability Accountability Act (HIPAA) regulations stipulate that each of us has the right to ask for a copy of our health records (http://www.hhs.gov/ocr/privacy/hipaa/understanding/consumers/consumer_rights.pdf), a scanned medical record is not a useful form of computable information. Without technology and systems in place that allow data from different sources, like your electronic medical data record and your FitBit device, to interoperate, meaningful data-sharing is nearly impossible. The end result today is countless small, siloed data sets locked away due to a lack of interoperable standards or by the terms spelled out in the original consent form.

The current privacy laws on health data also limit the ability of researchers to make correlations between an individual's genome and his or her medical health data. Although intended to protect people from harm, these laws that guarantee privacy of one's health data wind up blocking the possibility that good things could come from sharing health information and are fast becoming anachronistic in an internet age where sophisticated search algorithms can easily correlate clinical study DNA with publicly available information and re-identify individuals out of massive data[5].

So just putting our health data into an information commons of biomedical research is not sufficient. We need to make the data interoperable and to set common consent standards for its use; without these, the sheer quantity of health data that can now be produced will soon overwhelm us.

To learn what informed consent standards are best suited to an information commons, we need to pilot a whole spectrum of ways to manage privacy. Alongside our efforts sit the impressive work of the Personal Genome Project (http://www.personalgenomes.org/), the Vanderbilt Biobank Consent[6] and the scholarship of Latanya Sweeney[7].

We believe that a standard form of consent for the 21st century should be portable, reusable and honest regarding the risk-benefit of de-identification for studies that involve online data publication. Sage Bionetworks and John Wilbanks (chief commons officer at Sage Bionetworks and senior fellow in Entrepreneurship at the Ewing Marion Kauffman Foundation) are experimenting with a standardized way to provide a generic form of consent that is called portable legal consent (PLC; http://weconsent.us/informed-consent/). PLC makes it possible for data donors to carry their consent with them and to attach it to any piece of data they donate. Importantly, PLC is intentionally designed as an open standard to guarantee that all the data coming from different 'open consent' projects will be legally interoperable and usable as a common pool of information. Using PLC, anyone who signs up for a clinical trial, or has his genome read, should be able to share his genomic and health data easily, not just with that research group or company, but with all scientists who agree to follow a few simple rules about how they use the data.

Sage Bionetworks is already using PLC to see what happens when a small but coherent group of participants share their health data. The Self Contributed Cohort for Common Genomics Research (SCC-CGR) study asks participants to share their genomic, phenotypic, clinical and lifestyle data and has been well received by the research and patient advocacy communities. Those interested in signing up can do so using an institutional review board (IRB)-approved online tutorial (https://plc-cgr.weconsent.us/legalconsent/www/wizard/start.action). Sage is also in the process of implementing PLC on up to five separate studies focused on diabetes, chronic fatigue syndrome, Fanconi anemia, breast cancer and melanoma. And the University of California, San Francisco, has just written the PLC standard into one of their recently funded biobanks. Thus far, PLC is valid only in the United States, although Sage Bionetworks is looking at ways of adapting it to fit the legal frameworks of China and the European Union.

Synapse: an open compute platform

PLC makes it possible for citizens and patients to collect their own data or get them out of institutions and donate them to an information commons where they can be worked on all the time. In considering the necessary ingredients for a biomedical information commons, we have studied successful online examples like Wikipedia, Creative Commons and GitHub and noticed that each empowers their users to be creators, producers and distributors of information.

Similarly, we believe that a biology information commons that aggregates biomedical data and that leverages 21st century capabilities, like cloud computing and open social media, is precisely what is needed for data-driven precision medicine to take hold. In the past 20 years, biology has become an information-rich, data-driven discipline; the bottleneck is no longer at the point of affordable data production but now centered on our ability to transform data into meaningful information. Biology

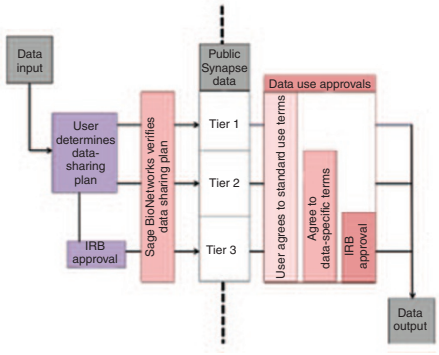

Figure 3 Map of Synapse data access governance to allow data sharing within a defined legal and ethical framework. Data governance within Synapse relies upon both policy and technology for its effective implementation. As data are placed into Synapse, their use are defined by the data contributor(s) and by IRB approval (the latter for human data). Sage Bionetworks confirms the data-sharing plan and assures that the data are hosted according to the plan. Researchers access the data in Synapse after agreeing to a terms-of-use policy that defines how that particular data set can be used.

desperately needs an open data repository equipped with tools that enable rapid learning.

To show how this can be done, Sage Bionetworks has, for the past two years, been building Synapse (https://synapse.sagebase.org/), an open compute platform that fosters transparent, reproducible collaborative research. In essence, Synapse is a computational geek's sandbox where data are continuously aggregated and productivity tools are in place so that integration, analysis and publication of data-intensive science occur in real time as the research is performed (**Fig. 2**).

Synapse is currently in beta release and, through partnerships with Google and Amazon, is leveraging the power of cloud computing to operate as an online resource for its scientific users. To establish a large interoperable data resource accessible to the public, Sage Bionetworks has invested in developing pipelines and workflows for the curation of large public data repositories, including over 12,000 data sets from The Cancer Genome Atlas (TCGA), Array Express and the Gene Expression Omnibus. Out of this curation pipeline comes publicly accessible, 'analysis ready' data that can be run anywhere, in the cloud, or on a user's personal laptop. To address the challenges of reporting clinically relevant computational data[8] in a way that is transparent and allows others in the field to reproduce the work, Synapse is embedded with provenance tools that produce formal records in real time of how data were processed by a researcher.

In keeping with the vision of Synapse as a potential biomedical information hub, social media and tools of attribution will soon be built into the Synapse environment to create a community of connected users who are recognized and rewarded (i.e., points and badges) for collaborative work. With all of this, Synapse will record and reward what people do, correlate it with their user profiles and allow them to share their work with anyone. Working with the tools currently available, Synapse's users have thus far generated more than 1 million computational models.

Because much of the data hosted on Synapse derive from patient samples, Sage Bionetworks has worked closely with an independent ethics advisory team (Sage Bionetworks Ethics Advisory Team; SB-EAT) to develop Synapse's data governance policies that allow data sharing within a defined legal and ethical framework. These policies balance the appropriate protection of human participants with the collaborative data sharing necessary to advance open science. Recognizing that different data sets come with different privacy concerns for their human participants, Synapse tailors data use restrictions to each data set using a tiered data-access system (**Fig. 3**). These tiers differ in the use restrictions, ranging from unrestricted access (tier 1) to IRB approval only (tier 3).

We and others are starting to see evidence that working in a shared space, such as Synapse, can help scientific teams find and correct analysis errors more quickly, get more people working on complex problems and combine analysis-ready data from more sources to answer scientific questions. For example, the Synapse platform has been recently selected by the TCGA pan-cancer consortium group (comprising groups from the Broad Institute, University of California Santa Cruz, the Institute for Systems Biology and Sage Bionetworks) for hosting all analysis ready data to facilitate the group's comparative analysis of TCGA genomic data. In addition, the Mount Sinai School of Medicine (New York, NY) has chosen Synapse to support groups working in Alzheimer's disease and diabetes. These collaborations constitute important first-use cases for Synapse that will help guide its further development.

The increasing complexity of modern biology is quickly rendering small self-contained approaches to research more and more untenable. Synapse showcases how a precision medicine information commons might operate and shifts current research practices in three ways. First, Synapse seeks to make results more reproducible by capturing intermediate analytical results and analysis provenance records as soon as a scientific team produces new work. Second, Synapse shifts work out of closed silos by providing researchers with the tools they need to publish their data analysis in an open online form, to track and document their own analysis progress, and even to join or track the progress of other researchers. Finally, Synapse shifts work from individuals or pre-defined small teams to open, evolving and scalable networks of distributed teams. Researchers who join in these networks will benefit; their own research ideas are enriched and evolve, and their opportunity to publish and receive funding will expand by virtue of their participation in a wide range of team projects.

Testing approaches and tools to galvanize sharing

In addition to needing new types of infrastructure and tools, we believe that a biomedical information commons also needs new approaches in place to drive the way research is performed and communicated. For instance, as researchers, we need to share information in real time and to function as part of a distributed and continually evolving team, but this is made difficult to impossible by the current reward structures that define career advancement by first-author publications and grant funding geared to principal investigators. Both the complexity of modern 'big data' biology and a commons approach to biomedical

Public Sage/DREAM Modeling Competition Leaderboard

ID	Name	Submitter	Status	Received	Train Score	Test Score	Detail
syn1417992	Attractor Metagenes Model 093009	317809	complete	2012-10-01 22:23:55	0.7502967496	0.7224532707	Output
syn1417956	Attractor Metagenes Model 093005	317809	complete	2012-10-01 22:23:54	0.7514339936	0.7219894559	Output
syn1417950	Attractor Metagenes Model 093004	317809	complete	2012-10-01 22:23:54	0.7509186799	0.7214792598	Output
syn1417990	Attractor Metagenes Model 093008	317809	complete	2012-10-01 22:23:54	0.7547248936	0.7192993306	Output
syn1417946	Attractor Metagenes Model 093003	317809	complete	2012-10-01 22:23:54	0.7240264125	0.7153414449	Output
syn1417893	WY093001	317809	complete	2012-10-01 22:24:05	0.7626456018	0.7148621697	Output
syn1417836	AT092857	962237	complete	2012-10-01 22:23:54	0.7510395121	0.7126203986	Output
syn1417942	Attractor Metagenes Model 093002	317809	complete	2012-10-01 22:23:54	0.7652692781	0.7122957283	Output
syn1417442	AT092753	962237	complete	2012-10-01 22:23:54	0.7506663539	0.7115845457	Output
syn1417446	AT092754	962237	complete	2012-10-01 22:23:54	0.7503620429	0.7115072432	Output
syn1417586	BH092852	962237	complete	2012-10-01 22:23:55	0.7511958832	0.7112134939	Output
syn1417366	BH092851	962237	complete	2012-10-01 22:23:55	0.7512385298	0.7109661261	Output
syn1417634	AT092856	962237	complete	2012-10-01 22:23:54	0.7508547100	0.7107496792	Output
syn1417473	BH092756	962237	complete	2012-10-01 22:23:55	0.7500941780	0.7104404694	Output
syn1418000	BasicModel3_0930_d07	342024	complete	2012-10-01 22:24:00	0.7632151310	0.7102085620	Output
syn1417994	BasicModel3_0930_d04	342024	complete	2012-10-01 22:24:00	0.7654967269	0.7101931015	Output
syn1417980	WarwickSystemsBiology (Sun Sep 30 20:20:44 2012)	362302	complete	2012-10-01 22:24:09	0.8020639997	0.7099611942	Output
syn1417440	BH092752	962237	complete	2012-10-01 22:23:55	0.7500670702	0.7098993522	Output
syn1418006	BasicModel3_0930_d09	342024	complete	2012-10-01 22:24:00	0.7657454990	0.7095746819	Output
syn1418129	BasicModel4_1001_n03	342024	complete	2012-10-01 22:24:01	0.7526671926	0.7094509980	Output
syn1417630	AT092855	962237	complete	2012-10-01 22:23:54	0.7499093759	0.7093736955	Output
syn1418002	BasicModel3_0930_d08	342024	complete	2012-10-01 22:24:00	0.7674655806	0.7093582350	Output
syn1417598	WarwickSystemsBiology (Fri Sep 28 19:45:01 2012)	362302	complete	2012-10-01 22:24:06	0.7974639458	0.7092500116	Output
syn1418018	BasicModel3_0930_D03	342024	complete	2012-10-01 22:23:59	0.7574827103	0.7092190906	Output
syn1418010	BasicModel3_0930_d10	342024	complete	2012-10-01 22:24:00	0.7619215159	0.7089098808	Output
syn1417570	WarwickSystemsBiology (Fri Sep 28 16:56:49 2012)	362302	complete	2012-10-01 22:24:06	0.7973644370	0.7088325783	Output
syn1417574	WarwickSystemsBiology (Fri Sep 28 17:16:08 2012)	362302	complete	2012-10-01 22:24:06	0.7976061013	0.7087088944	Output
syn1417465	AT092755	962237	complete	2012-10-01 22:23:54	0.7523544505	0.7085852105	Output

Figure 4 The real-time leaderboard used in the Sage Bionetworks/DREAM Breast Cancer Challenge. Participants develop and train their computational models on breast cancer patient data and then submit their developed models to Synapse where they are immediately scored (using held-back patient data). The real-time leaderboard lists the ranked scores for each computational model, the individual or team who submitted the model and a link to the source code for the submitted model that all can access.

research require that the current reward structures be inverted.

Sage Bionetworks is piloting several efforts to learn what motivates data sharing (computational challenges), patient involvement (BRIDGE) and rapid learning (clearScience). All of these approaches are designed to help identify the core principles needed to build an information commons where people come together to answer questions about disease and build multidimensional maps of health.

Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge (BCC). Sage's most successful attempt to date to align the incentive structures of biomedical science with more open and collaborative approaches has been the hosting of the BCC (Margolin, A.M *et al.* unpublished data). The BCC was a partnership with DREAM (dialog for reverse engineering assessments and methods), a network biology initiative led by Gustavo Stolovitzky (IBM; Armonk, NY, USA) that has run 24 successful computational challenges in systems biology in the past 7 years. By framing a scientific question and providing the clinical and genomic data to any interested participant (through access on Synapse), Sage and DREAM were able to motivate over 50 scientists or teams working around the world and using standardized virtual machines donated by Google to submit >1,700 predictive computational models of disease severity, all accessible and publically hosted on the Synapse platform. Participants' models were assessed by evaluating the concordance index between a model's predicted survival and the true survival information. The resulting model scores were immediately posted to a real-time leaderboard that grew in entry number throughout the three-month model-training period (**Fig. 4**). Participants' source code was available so that code sharing and collaborative model building could take place.

Prompted by the poor availability of potential clinical classifiers that can be successfully validated in patient cohorts other than the one used to construct them, Sage and DREAM determined the overall winner of the BCC by assessing participants' computational models against another breast cancer data set. For this final assessment, the Avon Foundation (New York) generously funded the creation of a never-before-released validation data set. The participant with the best performing model in the new data set was invited to publish an article about the winning model in *Science Translational Medicine*[10] and will present the model this month at Sage's 4th Annual Congress. To the best of our knowledge this is the first instance of a challenge-assisted peer review, in which performance metrics in a blinded validation test are the foremost critera for publication in a journal.

Structuring the BCC as an online competition proved to be enormously beneficial in motivating BCC participants to share code, interact on a forum (http://support.sagebase.org/sagebase) and form a cohesive community. The Challenge resulted in the development of predictive models that achieved a high level of performance across multiple blinded evaluations, including a novel validation cohort[11]. The BCC models also outscored both models based on standard clinicopathologic measures as well as a model based on the breast cancer 70-gene risk signature that is the basis of the clinically approved Mammaprint diagnostic marketed by Agendia (Amsterdam) and used to provide risk assessment for breast cancer survival[11]. Based on the success of BCC as well as DREAM's leadership to establish a growing interest and level of participation in computational Challenges, Sage Bionetworks and DREAM formally merged their efforts in early 2013. Working together, Sage and the DREAM team will leverage the Synapse platform and PLC to run computational Challenges that can potentially include newly generated data donated from patients and that seek to expand the number of diverse groups working on biomedical research in an open way.

BRIDGE. We believe that the most efficient and affordable approach to generate longitudinal multidimensional maps of health is to build a patient-centered system that positions citizens and patients as 'knowledge experts' who are empowered to collect and donate their data to open research. Toward this goal, Sage Bionetworks and Ashoka Changemakers (http://www.changemakers.com/main), with funding from the Robert Wood Johnson Foundation (Princeton, NJ, USA), have been piloting BRIDGE, a web-based platform intended to determine the operational requirements and value of citizen participation to both generate new data and provide energy and insights into disease research.

On BRIDGE, citizens, patients and researchers can form a community and use BRIDGE's tools to aggregate new research questions that matter most to patients and their families. Participants can then use PLC to collect their data and donate them to Synapse, where they are available to all for rapid learning and open challenges, such as the BCC. The insights and potential disease predictors resulting from the community research challenges and projects will be reported on BRIDGE and also become the basis for validation experiments.

Seven disease community groups (breast cancer, chronic fatigue syndrome, diabetes, Fanconi anemia, irritable bowel disorder, melanoma and Parkinson's disease) are either running or interested in running seed projects on BRIDGE. They all want to learn what it takes (i) to bring the open source movement to medical discoveries, (ii) to activate citizen-patients and (iii) to motivate the research community to share data and disease models. Each of these seed projects is designed to deliver something that could not have emerged using standard approaches. This could consist of a new approach to care, a new disease or treatment classifier, or a new insight that can be validated.

Mount Sinai's BRIDGE project on chronic fatigue syndrome (CFS) illustrates how BRIDGE's web-based community-building tools will be leveraged to complete the first-ever systems analysis of CFS. In this study, 300 participants (150 CFS cases and 150 healthy matched controls) will use BRIDGE's tools to form a tight-knit community and to provide regular personal narratives and an extensive amount of molecular data on themselves into Synapse. These data will be integrated in Synapse so that researchers can study CFS at a systems level and generate experimental hypotheses regarding the role that different components of the innate and adaptive immune system play in patients with CFS. The CFS patient community is already an incredibly activated group and we expect them to enthusiastically leverage all of BRIDGE's infrastructure and tools to collect and interact with their individual data and to start correlating their data with that of other CFS participants using BRIDGE. With the CFS data deposited into Synapse, a broader cross-section of citizens and researchers will be empowered to tackle CFS and move the field toward increased understanding and potential novel diagnostics and treatments.

clearScience. Besides piloting the Synapse platform and running computational Challenges to drive the way research is performed, Sage is also looking at new approaches to improve the currently low reproducibility of predictive computational models. We believe that this reproducibility problem could be vastly improved if computational biology researchers had the necessary tools and were offered incentives to use them to record the specific details of their code, data versions and computational environment.

Through funding from the Alfred P. Sloan Foundation (New York), Sage Bionetworks is collaborating with the editors of several scientific journals (*PLoS Computational Biology*, *Nature Genetics*, *Science Translational Medicine*

and *eLife*) to position Synapse as a platform that supports richer communication of data-intensive science. This effort, called clearScience, takes publications off the printed page (a 16th century technology), and reconfigures them using modern technologies that aim to eradicate the boundary between doing and communicating data-intensive science. clearScience allows narrative text to be read as HTML (hypertext markup language) and provides the resources and artifacts so that readers can recreate data analyses and critically access the analyst's assumptions, work and output. The platform even provides tools so that the data and outputs can be revised, updated and published by new analysts.

Sage is currently leveraging the open application programming interfaces of GitHub, Amazon Web Services and Synapse to integrate a beta version of clearScience into the Synapse platform where the growing user community can try it out and optimize its development. In addition, the first manuscripts to pilot the clearScience approach are currently being prepared and will be based on original research at Sage Bionetworks.

The next step: harnessing Metcalfe's law

Sage's team believes that biomedical research should be viewed as an "infinite game"[11] with an acknowledgment that there are no beginnings or ends to our understanding of complex diseases. We think that the rewards should be measured by new biological discoveries and new medicines that benefit patients, not by papers cited or patents filed. The tools and infrastructure that we and others are now building to develop the technology platforms capable of nurturing evolving data and models, the new informed consent standards returning control of health data to citizens, and the incentive structures in the DREAM Challenges and clearScience publication projects now afford us a powerful opportunity to work in a different way. If we are to move beyond the primacy of isolated teams generalizing from the particular to the general using hypothesis-driven science that has driven biomedical research in recent decades to Jim Gray's 'fourth paradigm'[12] of data-driven discovery, we will need a set of early adopters to be willing to tackle a few audaciously large problems like CFS by being part of a team and jointly working off of each other's insights.

One of the most exciting features of today's social media approaches that underlie Wikipedia, Flickr and the computational modeling efforts of EteRNA is that people have become components in a social distributed engine of idea generators. Bob Metcalfe, who co-invented the Ethernet, realized that when machines were visible to each other they could gain a power that required a critical mass but then scaled in utility in an quadratic fashion that has been called Metcalfe's law[11]. Metcalfe's law has been shown to apply to components that are compatible with each other, such as FAX machines and the Internet. This same power of compatibility can apply to people who are able to work on projects if they can share data and a common platform. Even though the genomic technologies and portable health devices have massively expanded the dimensions of new health information, we still need to develop enough compatibility in how these data and the models can be shared to harness Metcalfe's law if we are going to build a biomedical knowledge expert network sufficiently robust to bring about the promises of precision medicine.

ACKNOWLEDGMENTS

This article is intended to highlight topics of the Biomedical Information Commons that will be discussed at Sage Bionetworks' 4th Commons Congress (http://sagecongress.org/WP/2013agenda/) in San Francisco on April 19 and 20, 2013. Many will be streamed live on the web.

1. Hewitt, J. *et al. Health & Life Sciences, Oliver Wyman Point of View. Beyond The Shadow of a Drought: the Need For a New Mindset in Pharma R&D* (Oliver Wyman, New York, 2011).
2. Lyssenko V. *et al. N. Engl. J. Med.* **359**, 2220–2232 (2008).
3. National Research Council (US). *Committee on a Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (National Academies Press, Washington, DC, 2011).
4. Wruck, W. *Brief. Bioinform.* published online, doi: 10.1093/bib/bbs064 (9 October 2012).
5. Gymrek, M. *et al. Science* **339**, 321–324 (2013).
6. Roden, D.M. *et al. Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
7. Sweeney, L. *Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3* (Carnegie Mellon University, Pittsburgh, 2000).
8. Gentleman R. *Stat. Appl. Genet. Mol. Biol.* **4**, Article 2 (2005).
9. Carse, J. *Finite and Infinite Games* (The Free Press, New York, 1986).
10. Cheng, W. et al. Sci. Transl. Med. (in the press).
11. Shapiro, C. & Varian, H.R. *Information Rules* (Harvard Business Press, Cambridge, MA, 1999).
12. Gray, J. *Fourth Paradigm: Data-Intensive Scientific Discovery* (Hey, T., Tansley, S. & Tolle, K. eds.), (Microsoft Research, Seattle, WA, 2009).