Synapse Vision Document

Michael Kellen

The Problem

The past two decades have seen an amazing exponential growth in the availability of genetic and other important biomolecular data fueled by incredible advances in measurement technologies. These breakthroughs have resulted in an increasing amount of money spent on research by industry and research papers written by academia. However, despite a few examples of block-buster drugs, we have not witnessed a rapid improvement in the treatment of many significant human diseases. Indeed, the numbers of new drugs approved by the FDA has actually declined over this same period.

A fundamental reason improvement in clinical outcomes does not match the pace of biological data production is that the analysis and interpretation of this data remain largely an individual activity limited by the bandwidth of an individual scientist. Pharmaceutical R&D pipelines, and even many pre-commercial research programs, consist of a series of handoffs among individual scientists with different areas of expertise. In our experience, the ability to access, understand, reformat, and reuse data, analysis methods, or models of disease is a significant rate-limiting step in research progress, even within the confines of a single company or research institution. Additionally, much of the relevant data to answer a particular research question is spread among multiple public and private repositories. Because each pharmaceutical company and academic group protects their own data, the end result is enormous duplication of effort and missed opportunities across the industry as a whole.

Solution Strategy

Is there a better way to do research? Compare the situation in pharmaceuticals to the technology industry. In that industry Moore's Law, an exponential increase in basic compute power, is successfully translated into a wealth of new product launches every year. Here, some of the most widely-used software projects in the world are open source, including the Android operating system, the Apache web server, and the MySQL and PostgreSQL databases. Furthermore, the successes of these open-source projects have not killed off large corporations. Rather, by making some commonly-needed basic infrastructure available at little to no cost, open source software has lowered the barrier to entry to the industry as a whole, and seeded innovation as entire new businesses (Facebook, Twitter, etc.) spring up by innovating on top of this open base.

Additionally, in the software industry it is increasingly easy for large, distributed development teams to manage their development efforts using standardized infrastructure and tools. Currently, the trend is for software teams to outsource the hosting of code and supporting resources to organizations dedicated for this purpose (e.g. SourceForge, GitHub, Google Code, Atlassian). The tools provided by these sites (e.g. version control, bug tracking, wikis, automated build systems) let developers anywhere start a software development project and instantly access many supporting tools, and are used by open-source and professional software engineering teams alike. The success of many open-source development projects demonstrates that even highly distributed and decentralized teams can effectively collaborate on complex and large-scale projects given an appropriate collaboration framework.

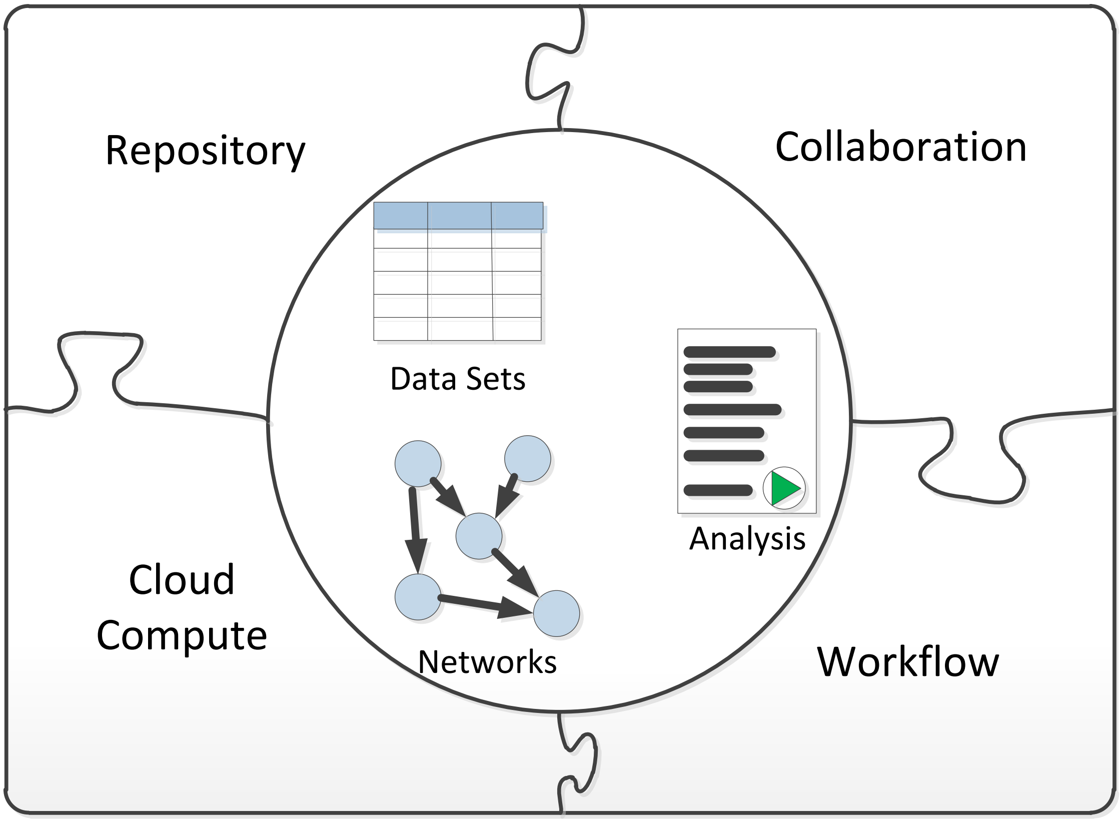
Obviously, there are some fundamental reasons why the pharmaceutical industry cannot simply copy the tech industry: the high cost and inaccessibility of experimental biology and the complexity of the regulatory environment make for a substantially different environment. However, when it comes to the analysis of large-scale biological data sets, particularly in areas of basic research not directly related to the performance of a particular proprietary compound, big Pharma could stand to act a bit more like Silicon Valley. If Netflix can put data on which people like which movies into the public domain to improve the methodology of their recommendation engine, why can we not do the same for (suitably deidentified and protected) human health data for a far more important purpose?

Sage Bionetworks' mission is to catalyze a cultural transition from the traditional single lab, single-company, and single-therapy R&D paradigm to a model based on broad precompetitive collaboration on the analysis of large scale data in medical sciences. If this were to happen it would benefit future patients through improved treatments to diseases. It would benefit society as a whole through reduced costs of health care. It could even benefit the pharmaceutical industry, which is struggling to replace the revenue lost as old medicines go off patent production, by seeding increased innovation in the sector. Sage Bionetworks is actively engaged with academic, industrial, government, and philanthropic collaborators in developing solutions to these issues.

The technology component of Sage’s solution strategy is Synapse, an informatics platform for open data-driven science. Synapse will provide support for Sage Bionetwork’s own research initiatives, but more importantly serve as a public resource for the broader scientific community. Although Sage does have some ability to create top-down pressure to drive people towards a platform (e.g. by influencing publishers’ or funders’ requirements), it is illustrating that open source code repositories do not grow by mandate. They grow because individual development teams are more effective when they let somebody else create, host, and maintain both this infrastructure and their own project code, and when development team members can easily build on each other’s work in a shared environment. Sage platform developers must aim to create a product useful enough that early-career scientists would actually choose to use it on their own volition, regardless of any top-down push from senior researchers.

Some of the basic pieces to create the technology infrastructure to support this mission already exist. A large number of repositories provide access to a wealth of biological data from a variety of studies (dbGaP, GEO, etc). A large number of tools are freely available that provide mechanisms to analyze common data formats (Bioconductor, GenePattern, etc). The leading computational biology groups can pull together multiple datasets and tools into sophisticated analysis pipelines that allow them to construct models that predict biological behavior, although this methodology is still at a relatively immature state.

However, the only way you can typically learn what these scientists has done is through a journal article that describes the work in a manner most often wholly inadequate to reproduce the analysis or apply it to a new project. What is missing is a place where biological data scientists can collaborate on these analysis projects in real-time and across organizational boundaries, a place where young investigators can actively learn methods from the leading scientists in the field by seeing exactly what was done, and a place where relevant data, tools, and models are brought into a single shared space with sufficient compute resources to support modern biological data analysis.



Catalyzing a transformation to truly collaborative research requires that the platform help scientists solve a series of problems that impede truly collaborative work today:

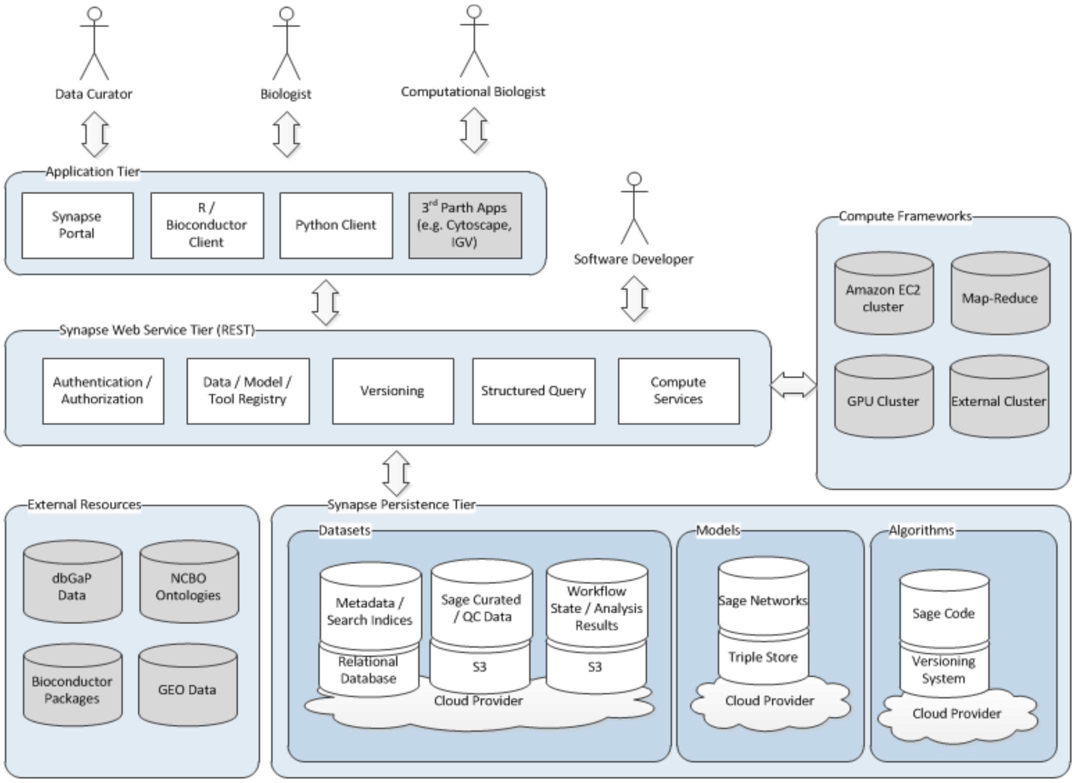
1. **Finding and using relevant data -** Frequently today scientists have difficulty just tracking down and gaining access to resources created by others, even within the confines of the same organization. Even when data is available, ~75% of the work of an analysis project is spent understanding data structures and preprocessing data for analysis. Synapse provides ways to not only make data discoverable, but push data to leverage common formats, controlled vocabularies, and annotation standards that allow data to be described and exposed to analysis unambiguously.
2. **Understanding analysis workflows -** Synapse accounts for the fact that much research is experimental and ad hoc in nature, and that hardened analysis methods only emerge over time. Tracking who has run what version of code on what version of the data in a project is the key to reproducible, and therefore improvable, research.
3. **Supporting genome-scale analysis** - Analyzing datasets with information on whole genomes is currently limited to those with access to large computational resources and significant IT support. Synapse makes cloud computing technologies accessible to scientists.
4. **Forming and maintaining productive collaborations** – When in doubt, scientists tend to start from scratch on a project rather than take work in an unknown state. The platform must help scientists track what work has already been done in a particular area and help create and sustain collaborations.

Software Architecture

To support the vision described above, a variety of types of users will need to interact with Synapse for different purposes. Initially the Synapse system will be focused on supporting the statistical or computational biologist power users in manipulating clinical and genomics data sets; over time support could grow for more biological or clinically focused researchers. Synapse leverages a web service-based architecture in which a common set of services is accessed via different sets of client applications to support a growing set of use cases over time.

One client application is the Synapse web portal. This is a "Web 2.0" environment for scientists to discover and share data, models, and analysis methods. The portal is organized around projects, which any scientist can create and invite collaborators to join. These online workspaces then serve as the glue to help teams of researchers collaborate to solve complex scientific analysis problems. Additional capabilities to visualize data and interact with disease models through the web portal will be developed over time.

We also expect our target users to be already proficient using data analysis tools, and to want to continue using those tools to work with data. Synapse's first analysis client will be an integration with the R / Bioconductor statistical package allowing users to track analyses performed with published packages or custom scripts to complete their work. Additional analysis platforms (e.g. Gene Pattern or Matlab) will get supported over time. All analysis platforms would interact with the Synapse system by calling the Synapse web service layer. This is a set of REST-based services providing support for annotating, querying, and updating data, analysis code, and models, and controlling access to these resources. These services will also allow tracking of the provenance of a multi-step analysis procedure, and executing analysis steps on cloud-based resources.



This extensible set of applications will need to access a variety of information including experimental data, network models, and algorithms. In some cases, this data could be geographically dispersed for a number of reasons (e.g. human genetic data can only be stored in dbGap). Even if data can be centralized, different types of data will likely need to be stored in different structures. Consequently, a repository services layer is necessary to insulate each application from the specifics of how each resource type is accessed and structured. This service provides a single location to handle annotation, indexing, auditing, security, and versioning.

Recently there has been an accelerating trend in the IT industry towards “cloud computing” environments in which large service providers (e.g. Amazon, Microsoft, and Google) provide on-demand access over the internet to shared pools of computing resources that can be provisioned, used, and released as needed. In addition to becoming an increasingly cost-effective strategy to provide compute and storage resources, cloud computing puts many basic IT management tasks (e.g. maintaining hardware, patching software, backing up data) into the hands of the cloud provider, thus keeping Sage focused on tasks that require scientific expertise. As a new development effort with little existing legacy, the Sage Platform will look to aggressively leverage and optimize its architecture to take full advantage of these existing, and coming services to provide large scale computation on demand to scientific applications.

At the same time, it will be important to avoid locking in users to particular cloud vendors, or to offer users the ability to conduct analysis locally when dictated by circumstance (e.g. requirements for particular hardware or to extract maximal value from servers already paid for). By taking advantage of abstraction layers like JClouds, Eucalyptus or Nimbus that provide a single interface to local or cloud-based resources, the platform will allow collaborators to move computation and/or data among environments as needed. Note that the volumes of data preclude moving data on demand at analysis time. Rather, each site where computation may be located will need a local copy of the data, and the repository services will need to insure that requests for data resolve locally when appropriate and synchronize data sets in the background; see also the Cloud Computing Requirements document on the intranet.

Sage Web Portal

The Sage Web Portal is envisioned as one of likely many client applications that use Sage Service. The Sage Portal will be a “Web 2.0”[1] environment for end user scientists to interact and share data, models, and analysis methods, both in the context of specific research projects, and broadly across otherwise disparate projects. General web-based collaboration / project management functionality is available from many large software vendors, and we expect to integrate with existing products in this area, see Project Collaboration Requirements. The Sage Web Portal should provide:

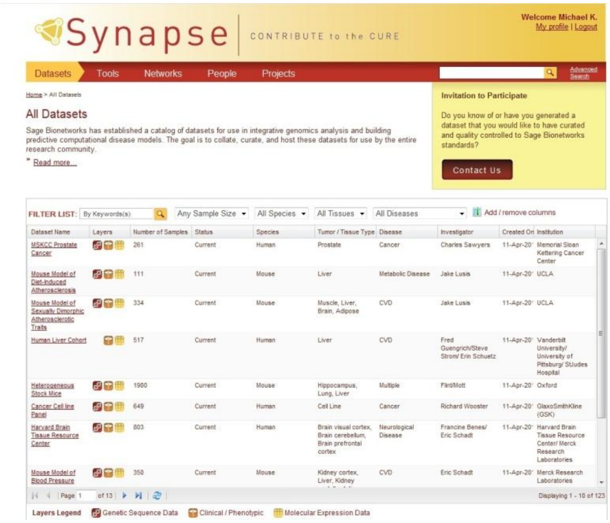
- Ways to easily search and navigate through content relevant to their research interests. This will includes a combination of browsing, free text search, and structured queries to find relevant data, tools and disease models.
- Ways for researchers to form projects to organize and track data analysis projects. Note that the portal is not intended to be the analysis tool itself, but rather to surface the provenance of each step in an analytical pipeline so that it is understandable to the rest of the project team.

[1] I know, one of the most overloaded and useless marketing terms ever coined. For me, Web 1.0 means one set of people publishing content on the web using some publishing tool, and other people reading the content using a browser. Web 2.0 means end users, using the browser itself, contribute to the content seen by other users.

- Curation tools to allow users to annotate and organize data. This will start as featured directed towards Sage's dedicated curation team, but should evolve to support allowing the community to curate data.
- Facilities for connecting people with overlapping research interests. The web portal is not intended to be a social networking site; however integrations with social networking sites like Linked-in or Facebook could be leveraged to help people connect.
- A framework for integrating and organizing the UI around a variety of tools added into the framework, conceptually similar to the way Facebook allows external developers to embed "apps" within Facebook's pages that access Facebook data. This would allow integration with a variety of analysis and visualization

Users will require an easy way to visualize networks stored in Sage through a client application that accesses Sage Platform network data. Export to / integration of Sage services with Cytoscape or other off-line tools is one option, but requires users to install and learn to use that particular tool. It could be useful to integrate a web-based network viewer that could provide basic network visualization to a broad group of users. This would not be a replacement for an expert-user tool, but a way to get very basic functionality to the masses (and good eye candy for fund-raising). There has recently been a release of a Cytoscape web version that would a first choice for this purpose.

A common follow-up activity to data analysis is mining the existing literature and a variety of public databases for more information on genes of interested uncovered by the analysis. In this area, there are a large number of potential commercial and open source solutions available to researchers including NextBio, Ingenuity, tranSMART, Gene Atlas etc. Replicating this functionality is outside the scope of Synapse. Synapse's core focus is supporting data analysis, and we will look to integrate with one or more of these sorts of platforms for follow up knowledge management and data mining.



All Datasets

Sage Bionetworks has established a catalog of datasets for use in integrative genomics analysis and building predictive computational disease models. The goal is to collate, curate, and host these datasets for use by the entire research community.

» Read more...

Invitation to Participate

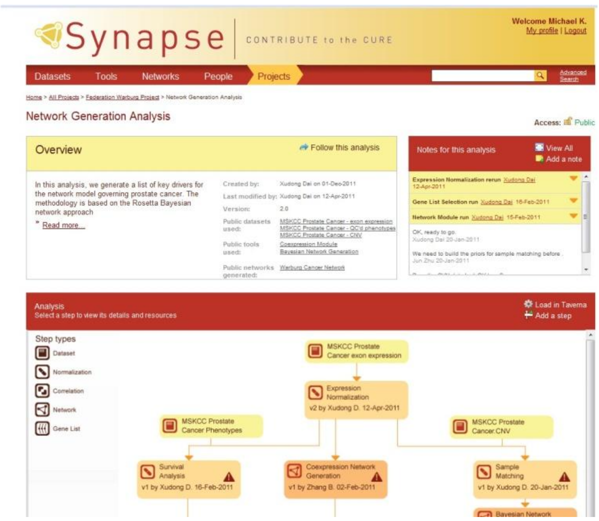
Do you know of or have you generated a dataset that you would like to have curated and quality controlled to Sage Bionetworks standards?

Contact Us

Dataset Name	Layers	Number of Samples	Status	Species	Tumor / Tissue Type	Disease	Investigator	Created On	Institution
MSKCC Prostate Cancer	🟥🟧🟨	261	Current	Human	Prostate	Cancer	Charles Sawyers	11-Apr-20'	Memorial Sloan Kettering Cancer Center
Mouse Model of Diet-Induced Atherosclerosis	🟥🟧🟨	111	Current	Mouse	Liver	Metabolic Disease	Jake Lusis	11-Apr-20'	UCLA
Mouse Model of Sexually Dimorphic Atherosclerotic Traits	🟥🟧🟨	334	Current	Mouse	Muscle, Liver, Brain, Adipose	CVD	Jake Lusis	11-Apr-20'	UCLA
Human Liver Cohort	🟧🟨	517	Current	Human	Liver	CVD	Fred Guengrich/Steve Strom/ Erin Schuetz	11-Apr-20'	Vanderbilt University/ University of Pittsburg/ StJudes Hospital
Heterogeneous Stock Mice	🟥🟧🟨	1900	Current	Mouse	Hippocampus, Lung, Liver	Multiple	Flint/Mott	11-Apr-20'	Oxford
Cancer Cell line Panel	🟥🟧🟨	649	Current	Human	Cell Line	Cancer	Richard Wooster	11-Apr-20'	GlaxoSmithKline (GSK)
Harvard Brain Tissue Resource Center	🟥🟧🟨	803	Current	Human	Brain visual cortex, Brain cerebellum, Brain prefrontal cortex	Neurological Disease	Francine Benes/ Eric Schadt	11-Apr-20'	Harvard Brain Tissue Resource Center/ Merck Research Laboratories
Mouse Model of Blood Pressure	🟥🟧🟨	350	Current	Mouse	Kidney cortex, Liver, Kidney	CVD	Eric Schadt	11-Apr-20'	Merck Research Laboratories

Displaying 1 - 10 of 123

Layers Legend 🟥 Genetic Sequence Data 🟧 Clinical / Phenotypic 🟨 Molecular Expression Data



Network Generation Analysis

Overview

In this analysis, we generate a list of key drivers for the network model governing prostate cancer. The methodology is based on the Rosetta Bayesian network approach

» Read more...

Created by: Xudong Dai on 01-Dec-2011
Last modified by: Xudong Dai on 12-Apr-2011
Version: 2.0
Public datasets used: MSKCC Prostate Cancer - exon expression
MSKCC Prostate Cancer - QC'd phenotypes
MSKCC Prostate Cancer - CNV
Public tools used: Coexpression Module
Bayesian Network Generation
Public networks generated: Warburg Cancer Network

Notes for this analysis

Expression Normalization rerun Xudong Dai 12-Apr-2011
Gene List Selection run Xudong Dai 16-Feb-2011
Network Module run Xudong Dai 15-Feb-2011

OK, ready to go.
Xudong Dai 20-Jan-2011

We need to build the priors for sample matching before .
Jun Zhu 20-Jan-2011

Analysis
Select a step to view its details and resources

Current designs of the Synapse web portal showing interfaces for: Browsing hosted datasets in Synapse (left); and Viewing a step in an analysis pipeline (right).

Sage Repository Services

The variety of resource types managed by the Sage platform will require their own dedicated storage: analysis code will need to be in a software versioning system, raw experimental data will be stored as flat files, analyzed data could be in a relational database, and network data could be in a triple store. Data may be additionally fragmented geographically due to a variety of non-technical requirements (e.g. dbGaP requires that human genetic data be shared via their own servers). Due to the pace of change of scientific data and requirements, there is a need to insulate calling applications from the specific details of how resources are discovered, identified and accessed. The Sage repository will provide a generic API for working with resources which may be biological data, networks, or algorithms via a URL following linked data principles[2]. The benefit of this approach is the reuse of these services across a number of potentially growing classes of resources. This layer provides a single place for a variety of general-purpose platform features:

- **Annotation** – Facilities for managing resource metadata associated with Sage resources that describe their structure and context. This includes leveraging controlled vocabularies or ontologies to ensure consistency across different resources. Where appropriate, emerging standards like CDISC for clinical data will be leveraged.

- **Search / Indexing -** Both structured and unstructured federated query mechanisms to find resources via indexes created by this layer[3].

- **Auditing** – A recording of the history of who did what to produce a particular resource, resulting in a high level history / work flow for projects run on the platform.

- **Security** – Resource level control on user level access and guest level access that can evolve over time to reflect the changing nature of resource availability with project life cycle. This layer will leverage emerging standards (e.g. Open ID) to manage access to platform resources.

- **Versioning** – Object level version history, with relationships between resources tracked at the version level. There will also need to be a distinction between published resources ready to be reused and work in progress.

Sage may not be the ultimate storage location for many datasets. Instead, it is desirable that certain datasets or parts of datasets live online in other public repositories where appropriate (e.g. NCBI GEO for gene expression data and dbGaP for clinical / genetic data), and that Sage use linked data principles to make data sets from these repositories available when needed[4].

[2] http://linkeddata.org/

[3] Although a counter-argument could be made that just "Google searching" free text is good enough, talk to Atul Butte on this point.

[4] This statement should not be taken to mean that Sage can not cache a *copy* of a remote dataset for performance reasons if needed, just that the platform must track who is the single point of truth for the data.

Sage Analysis Services

An emerging vision of cloud-computing is that data and the software that operates on it will mostly live on remote servers, accessed by people through a variety of mechanisms[5]. Instead of scientists moving large amounts of data around to use with local software applications running on local compute resources, Synapse should encourage scientists to upload analysis methods / tools, and run analysis on Synapse. There are a couple reasons for this: one is that the economics of the cloud model are becoming increasingly attractive and it will become increasingly rare for research teams to manage their own custom Linux cluster just so their scientists can do high performance computing[6]. An even more important reason is that the fundamental goal of the Sage Platform is to make sharing scientific data really happen. This won't happen by magic, the software must be designed with a "social interface"[7] that strongly encourages people to reuse other people's work ***and then open up the results of their work for others to reuse***. Scientists in general like to gain access to other people's work. Getting a scientist to go to a Sage Platform that contains a dataset (or analysis tool, or network model) that he already wants to work with will be about as hard as knocking a drunk off a tightrope with a baseball bat[8]. However, if that scientist chooses to happily download that dataset, run local analysis, and write his Nature paper, there will be a very strong chance that nobody will go to Synapse because nothing is in Synapse.

Of course running an analysis on Sage must be as easy and flexible as running analysis on a local machine, or it won't happen. The Sage analysis services must provide:

- **Data import / export** into commonly used analysis environments. A scientist should be able to work in R (or Perl, Python, Matlab, other specialty tool etc.) and load Sage data via a single call passing in the URI of the dataset.
- **Easily scale analysis –** mechanisms like increasing the size / numbers of VMs are an option, as well as providing analysis written for parallelization frameworks like MapReduce.
- **Capabilities for managing publication of results –** because there is a lot of experimenting and people probably don't want their colleagues looking at intermediate steps of debugging a new analysis routine. However, it's also pretty important that scientists don't just keep their own work private indefinitely while they go off to write their papers and grants.
- **Ways to distribute cost –** Use of cloud services opens a number of ways in which the revenue needed to maintain the platform would grow with increasing use. Ideally funding can be found to provide free access to services to the scientific community, but there will need to be some mechanism for tracking individual's usage of the system, and probably charging at least heavy users for their loads. This applies to compute cycles, but may actually be even more critical to consider in the context of the storage / IO required to maintain large datasets or network models. Amazon, has Flexible Payment and DevPay services that are designed to allow other organizations to build applications on top of Amazon web services, and bill users for their usage[9].

[5] In one example, Tim O'Reilly discusses his cloud computing vision. And yes, this is a bit consumer-focused and intentionally forward looking and it might not be reasonable to expect a high-throughput biologics platform to work the same way as your smart phone. But then again, we must heed O'Reilly's advice to build software for the world as it will exist in 5 years.

[6] Translational bioinformatics in the cloud: an affordable alternative Dudley et al, Genome Medicine 2010, 2:51

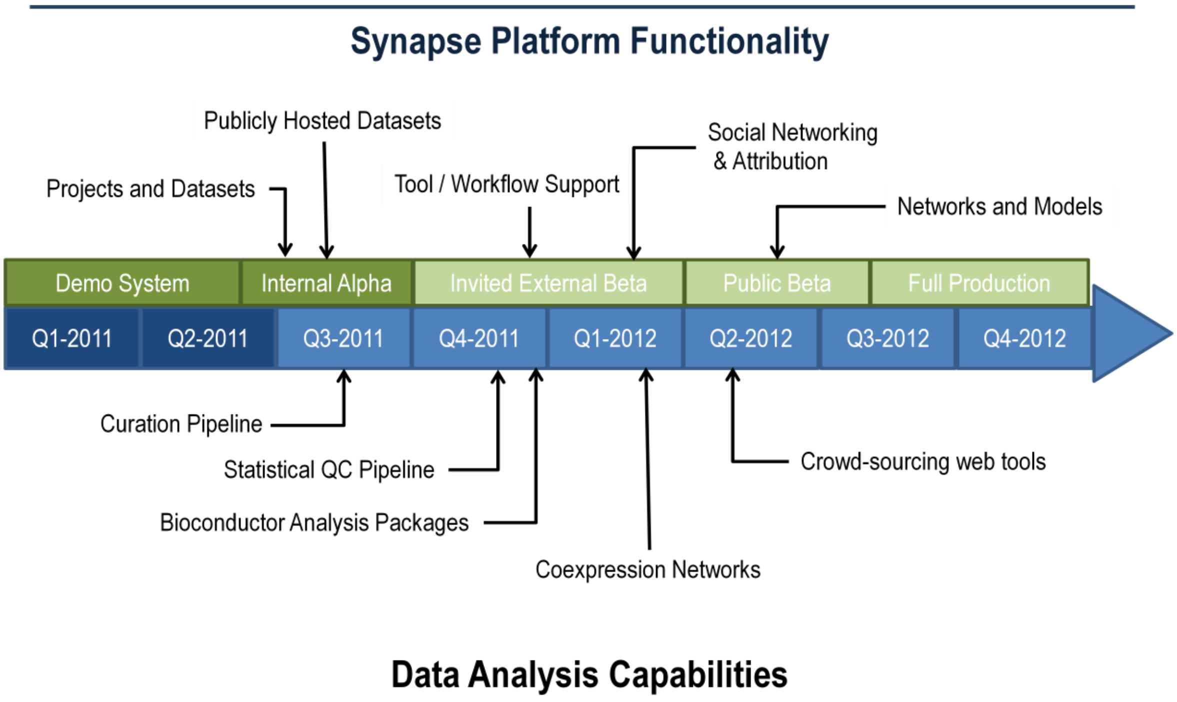
[7] http://www.joelonsoftware.com/articles/NotJustUsability.html

[8] I know. But now you're paying enough attention to even read the footnotes.

[9] These organizations are free to charge any price, or none at all, for their "value-add" services

Release Plan

The Sage platform team is now entering internal alpha testing and further public development after that, leading to a public beta by April 2012.


Synapse Platform Functionality
Publicly Hosted Datasets
Projects and Datasets
Tool / Workflow Support
Social Networking & Attribution
Networks and Models
Demo System
Internal Alpha
Invited External Beta
Public Beta
Full Production
Q1-2011
Q2-2011
Q3-2011
Q4-2011
Q1-2012
Q2-2012
Q3-2012
Q4-2012
Curation Pipeline
Statistical QC Pipeline
Bioconductor Analysis Packages
Coexpression Networks
Crowd-sourcing web tools
Data Analysis Capabilities