Synapse Demo Script

*9/6/2011*

This script is for use with the Sage demo.

***Bold italicized text*** indicates that you should be pointing or clicking on something.

[*Technical background notes*] should probably be skipped unless a question comes from the audience, or the presentation is to a technical, rather than scientific audience.

Live Synapse demo is online at <http://synapse.sagebase.org/Portal.html#Home:d>. Note that Home:d puts a bit in your cookie that puts the stack into demo mode, which means that the projects and their contents are hard coded to match the old static demo. <http://synapse.sagebase.org/Portal.html#Home:nd> puts you into normal mode where you will see Sage projects in progress. To give the demo you should have edit and/or admin access to the MSKCC data set.

# Introduction

General introduction might include statements about how pharmaceutical industry is struggling to bring new, innovative medicines to market and new approaches to biology are needed. Background on why Sage is taking the approach that it is towards open data and reproducible research. Give the 4 main themes of Synapse:

* Compute space for working with genomics data. Tools must capture work as it is actually performed, capturing at point of publication fundamentally does not work for real reusability.
* Ability to publish data, tools, models for public use; Sage committed to actively working to populate the system with a variety of content.
* Collaborative workspaces: increasing amounts of work across organizational boundries
* Cloud computing: rent compute power and data storage for both economic reasons and ease of sharing.

There is a Synapse Intro PowerPoint deck that you can use to go through this. But remember **keep it short** and get to the demo.]

# Home Page

Today I’m going to introduce you to Synapse, an online community and platform for data-driven research that is being developed to support the Commons. Through the rest of this demonstration, I’m going to take the role of a new scientist coming to Synapse to see what I can learn about active research in my field.

We’re now looking at the home page for the Synapse web portal, which is likely the first point of contact I would have with the platform as a first-time user. The ***main menu bar and intro section*** should give you a sense of the breath of functionality we’re looking to pull together. On Synapse, I have access to all the public data sets hosted by Sage. This includes data we’ve curated ourselves to seed the platform with generally available data, as well as any data sets made public from any of the projects run on Synapse.

But Synapse is designed to be more than a data repository. It is meant to be a place where scientists come to actively do analysis on datasets in a way that is more reproducible than traditional methods. Thus I also have the ability to access and run tools and network models created by other scientists, and string those tools into analytical workflows. These analyses may access data hosted by Sage, or data hosted by another organization. Another important point about our vision for the platform is that new data analysis methods will increasingly require scientists to move their analysis to large scale data-warehouses. So, I’d like to point out that everything you will see in today’s talk is running on top of Amazon Web Services to facilitate bringing cloud-compute capabilities to the analysis.

Finally, we also feel that there is a growing transition in biological data analysis towards team-based approaches, where scientists can build off each-other’s work. Reproducibility and versioning of data, tools, and models is one necessary function to support that vision, but also crucial are the social issues around how groups of people collaborate. Thus the final focus area for Synapse is on the people who use the system, and the projects and collaborations they can create on the platform.

***If I look around the home page***, I see evidence of this last focus area through the highlighted projects and people. I’d like to especially thank some of our key ***Synapse Contributors***, who are all real people who have already made a real significant contribution to the Commons by helping us release some datasets before we even have a rudimentary platform in place. So if you see them around the Congress please let them know much you appreciate their work. If I’m curious, I can get to the ***complete list of contributors*** working on the platform. If I recognize a colleague I might ***look at his profile*** to see what he’s working on. Basic information about his work is synched from Linked-In or another social networking tool. He doesn’t need to keep a separate profile at Sage up to date. I can also see the Sage projects and datasets he is following (we’ll go more into that later). Most of this profile information is optional, but we hope people will use the site to connect with colleagues and potentially new collaborators. We want to encourage some of this sort of behavior because science works off of peer review and recognition of scientists. [Note: *Users will need to provide their name and a valid email to access Sage data and make any contributions (from comments to starting a new project). However, they wouldn’t be forced to share detailed profile information.*]

I might contact Charles later, but for now I’m going to ***go looking to see what data Sage is hosting***.

# All Datasets

Now I’m interested to start ***browsing through data sets***. This page is where I can go “shopping” for data hosted by Sage. Here I see a list of datasets organized in a way to allow me to find the data that’s interesting to me. The layers column is a high level indication that expression, genetic, or clinical data of some sort is present for the dataset. I can see a number of additional fields that contain standardized annotations of the data. Here, I can ***click on a column header*** to sort by that column, for example to sort datasets by the total number of samples in the study. I can ***filter a column*** by selecting values from one or more drop-down filters, for example to only see Human datasets greater than 100 patients with data taken from prostate. [*Note that searching for datasets related to brain tissue should return me datasets marked as coming from cerebrum or cerebellum since those are regions of the brain. Not operational yet, but worth pointing out. In order to make this sort of functionality work, there will have to be some tie-ins to controlled vocabularies / ontologies during the curation process to standardize terminology across data sets. Also, this allows more sophisticated queries.*]

Or, I might do a broader ***free text search across all datasets*** (not implemented, just point out the functionality).

*Optional*: I’m not sure these are all the annotations I’d want to see, so I’m going to ***click to show additional columns***. Here, I can control what information is shown in the table, for example I might want to see how many globally coherent samples (or samples for which all layers are present) are in each dataset rather than the total number of samples. As an aside, the platform will be built in a way that it will be easy to change what annotation fields are captured by our curation process. The mockups are showing our best guess as to what people want to see here, but we aren’t locking ourselves into a decision here.

[*Note*: *We are currently have implemented REST-based web services to manage study and layer metadata, backed by a relational database. persistence mechanism for metadata; the goal is to not have schema changes as the fields change, which they certainly will over time*.]

# MSKCC Dataset

I am interested in Charles Sawyer’s data set, so I ***follow a link*** to a page of information on that data set. Here I can see a more complete description of the dataset, along with the complete set of annotations on that data. [*Note: Annotations are simple key value pairs where the value is strongly typed. We are actively working with the NCBO to connect Synapse annotations to existing ontologies where appropriate. Datasets and other Synapse resources are queriable via a REST interface, which takes a SQL-like string to search against the database. The back-end storage is an Amazon RDS MySQL instance.*] If you ***look in the lower-left hand corner of the page*** you can see exactly what sort of data is contained within this dataset, and how many samples we have for each layer.

# Curated Clinical Data Layer

I’m interested in the Curated Clinical data, so I’m going to ***click on that link***. Here I immediately see a preview of the data. I can ***pan left and right*** to view the first few rows of all traits and covariates in this layer. If I ***hover over a column header***, I will get a popup with a more complete description of that trait / covariate.

The annotations of the layer include some useful information, like the name of the person who performed the QC analysis, and the script used to do the QC. I’d be able to inspect the curated data and curation methodology if I wanted to check the QC work.

We also are working to embed curation tools directly in the Synapse web application framework. Our immediate goal is to make our own internal curation effort more efficient. Our long-term goal is to create a system whereby the community can collaboratively work to clean up annotations and data following a Wikipedia-type model. [Note: *this is another point of integration with NCBO*.] If I ***chose phenotype editor*** from the admin menu, I get a dialog allowing me to work with the data and help standardize its structure. Using this tool, I can set up column definitions that describe what each clinical variable is precisely… Columns can contain Numbers, text, enumerations, or values from ontologies for example. This is also where curators can set units on variables and longer free-text descriptions of the columns. Over time, we’d like to build up reusable column definitions that can be mapped to more than one dataset to further streamline curation. For now, I’ll just ***cancel*** out of this mode.

From here, I might ***download the data*** if I though it were useful. All users must be logged in and agree to terms of use to access Sage hosted data. In this case, I happen to be an administrator of the project that hosts the dataset, so I can ***edit the sharing settings*** to control access to the data. Here I can see that the public already has read access to the data, and some members of the Sage staff have edit and admin privileges on the data. [Note: *In most cases data access is controlled at the project level, but we have implemented fine-grained access control on data.]* For most data hosted by Sage this just requires a Synapse account. For some types of restricted health data like whole genome SNPs, we will have a tiered access program requiring a formal application to access restricted data. While the terms of access might be initially cumbersome, Sage as an organization is working to streamline the process as much as possible, and work on the policy side to improve the ability of scientists to share data for research purposes. [*Compliance note: Sage intends to eventually make public who is accessing what datasets for what purposes. It might be worth eventually making this explicit in the demo to get feedback*.]

I’m satisfied that this data is useful to me. However, I want to see who else is already working with it, so I’m going to return to the main data set page by ***clicking on the breadcrumb***.

# MSKCC Dataset

However, I’m not really interested in the data until Sage completes the QC of the genetic and expression data. So, I’m going to ***follow the dataset***. In my case, I’m interested in hearing from Sage anytime the dataset has a new layer or has new samples published, is used in a new project, or is referenced by a new publication.

If I ***view my profile***, I can see all the datasets, projects, and other platform resources I’m following. [*At some point, this information could be used to help introduce people who have similar research interests, although this functionality isn’t mocked up yet.*] However, I don’t really want to be bothered every time there is a comment on the dataset. That seems like overkill to me. Now that I’m satisfied with my settings, I’m going to go ***back*** to viewing the dataset.

Moving over from the dataset layers summary, there is an area where people can comment on the dataset itself. This could be linked into some email / RSS / other notification mechanism. I can also see a list of projects actively using the dataset to do research. I’m going to ***follow the link*** to the Federation Warburg project as I think I remember my friend Xudong is working on that one, and it’s using a dataset that I’m interested in.

# Federation Warburg Project

Here I am looking at a high level summary of the Federation project on the Warburg effect in cancer. I can quickly see who’s working on the project and what they are doing. In this case, we have three scientists from Sage Bionetworks, Stanford, and Columbia leading the project, and I could poke around a bit here and dig into a bit more about what they’re doing by checking out their list of publications or Google App site if they’ve posted anything there. [*Google Apps just happens to be an example of the general collaboration tools with which the platform could integrate. The platform is not providing any of this functionality directly. Instead we will just integrate with whatever general-purpose collaboration tool the team wants to use*.]

Notice in ***the upper right corner*** the privacy settings for this particular project. The project is open to the public for viewing. Of course, the existence of these settings implies that the project could have been restricted to some list of pre-defined collaborators. At this point, it’s probably obvious how browsing through an example of an analysis performed by a leading group might be of value to me as a new graduate student or post doc looking to start my career. But, what would motivate the actual project leads to open up the project? Wouldn’t it be in their best interests to gain access to Sage-hosted data as quickly as possible and then move on to conduct their work in private until after publication? In fact, plenty of target users who’ve seen previous versions of this demo have said that is exactly what they’re looking for out of the platform so it’s not a trivial question. We’ll come back to that question towards the end of the demo.

For now, I’m interested in seeing what more I can learn about the analyses conducted by this project team, so I’m going to look ***at the lower left hand corner*** of the screen where I see three analyses in progress, led by different team members. I’m going to ***follow the link*** to view the network generation analysis conducted by Xudong of Sage.

# Sage Network Generation Analysis

Here, I’ve got a quick overview section giving me more detail about what was done, but I’m going to pretty quickly ***focus on the graphical view*** of the analysis. I can see the first thing Xudong did was to normalize the expression data. The normalized data was then fed into 3 different analysis steps: First, it was combined with the clinical data to conduct a survival analysis. Second, it was used to generate a coexpression network with which Xudong got some help from Bin. Finally, it was combined with the copy number data in a sample matching analysis. Looking through the rest of the analysis, I can see how Jun helped generate a Bayesian network, after which Xudong annotated the network with the coexpression modules, ranked the modules using the results of the survival analysis, and then came up with a list of gene candidates of interest. This is a nice summary, but so far not much more than I might see in the methods section of a traditional journal article. The real value comes in when I want to learn more about a particular step in this analysis, which I can do by ***clicking on the expression normalization***.

# Analysis Steps

On this step, I can see more details about the normalization step. In particular, I can ***see the list of inputs*** which is a link back to the publicly hosted expression data. I can see the ***outputs which is a link*** to the table of intermediate results created by this analysis. And, I can get the ***R script*** used to generate the outputs from the inputs. If Xudong was really careful, he might have even used a technology like SWEAVE to create a nice human-readable vignette of his work, which would ***look something like this.pdf***. [Use Justin’s example or dummy something up?] I can also see that this is the second time Xudong has run this particular step of the analysis. ***Looking at the notes section***, I can see that Brig had some suggestion on how to better remove batch effects in this dataset, and gave Xudong a bit of help with this step. ***Looking more closely at the rest of the analysis pipeline***, I can see that Xudong still needs to rerun the rest of his pipeline to regenerate his results based on this change. That’s a useful thing for me to keep in mind as I look over his work.

Similarly, to understand how the coexpression network got built, I could ***click on that step*** in the workflow and again see the inputs, outputs, and analysis script for that step in the larger workflow. In this case the output of the analysis is a network rather than a table of data, and if ***I click to see the output***, I’ll get a view of a network. Here, we get a view of the network created during Xudong’s analysis, which I could interact with in the web client. I can see the group has published the network out for other groups to reuse in other projects, which is a nice act of either altruism or marketing on their part. I can actually see another group has reused this model in another project. From this point, I could load the network into a tool like Cytoscape for more detailed work. Instead I’ll return to the analysis and work through it to get to the final list of genes produced.

This would let me completely reproduce Xudong’s analysis, which would be the first step I would probably take if I wanted to adapt his approach to my own research.

# R Command Line

At this point, those of you in the audience who are more on the biologist side may be following along, but the more hard-core computational types may be thinking, “That looks nice and all, but all I really want it programmatic access to the data, not all this project tracking stuff.”

So, now I want to move into the role of someone who is actually going to contribute to the project. I’m going to switch over to the R Command line interface, which is where most of Sage’s analysts do their real work. All facilities shown in the web UI for querying and access Sage data are constructed as a stand-alone web application accessing a set of REST services. Those services can be accessed simply in a variety of clients and programming language. Here, we’ll use an R package to query the Sage repository, find this same dataset, and load a layer into R for analysis, all without a nice pretty GUI.

First I’ll make a call to find all datasets where the species is human and the disease area is prostate cancer. This gives me a list of dataset identifiers that match that search criteria. I can pull the metadata from the repository for a given dataset, including an enumeration of the layers contained in the dataset. I then can pull the annotations on the layer itself, and finally load the layer as an object in R.

Once I’ve gotten my input data loaded into R, I can work with it using R in any way I want. The methodology followed by our scientists is to create a script that takes the analysis from a defined starting point to some final state. Once this is done, the script and the analysis results can be checked back into the repository, as new steps in a particular workflow. Of course, I could just as easily get the normalized data, or a network, or any other intermediate step of the analysis if I were a project member. The platform is tracking who checks what in, what version it is, supporting documentation, all of which would be updated at the web client level when I ***returned to my browser***.

# Gene Expression Search

From the results of the prior analysis, I notice that the ROR2 gene was highlighted by the Federation analysis. I’d like to know if there are any other datasets hosted by Sage where this gene is differentially expressed. I’m going to search across the site for resources that match this gene. Note how ***as I type*** the application suggests scientifically relevant terms including gene names for me to select. I’ll pick the ROR2 gene and ***hit search***. [*Note I could have also done a search using the signature defined by this list as my search criteria.*]

On the search results page I get a list of all Sage-hosted datasets where gene expression is significantly up or down regulated between two conditions in the study. Sage has curated some representative subsets of GEO data, which are shown here. If I’d like to learn more about one of these studies I’ll ***click on it*** and go to the dataset view for that study.

Hopefully, by this point I’ve at least convinced you that if you are a computational biologist there are some benefits to organizing a project in this sort of world. In particular, if you are working on a distributed project with team members spread across the country this sort of approach might really help your team collaborate more effectively. And, if you could see the work others had done in this sort of environment, you might be able to learn far more than you ever could by reading a traditional journal article.

[Note: *At this point I usually go back to slide deck and talk about road map and timelines for releases*.]

# Future Tweaks

* The initial click to Charles Sawyers is a bit awkward, people want to start with the data. Makes more sense to get interested in collaborating with people after you find something interested they have worked on. Link to Sawyers off dataset page
  + Make sure link to Sawyers profile from MSKCC dataset page works.
  + Make sure link from Sawyers profile back to MSKCC dataset page works.
  + Rename creator -> Contributor
  + Bonus: Add Sawyer’s picture along with his name
  + Bonus: Add Matt Furia and Solly Siebers as curators
* Bonus: Project and analysis pages: Add pictures of leaders like Xudong
* Expression Normalization step: needs a version field in the Overview, text description of step should describe why Xudong is rerunning the step
* Search box should work on dataset pages so we can find the dataset another way early and not have to circle back to the project again at the end of the talk.