The Cognitive Hourglass
in Healthcare and Life Sciences R&D

exaptive

Executive Summary

Better utilization of the large, rapidly growing, and often dissimilar data sets in healthcare and life sciences can provide insights into many unanswered questions and address many of the challenges faced by research and development organizations today. Yet many organizations are still extracting only a small fraction of the investment they have made in data infrastructure and of the value that is hidden away in their data.

Why?

One reason is that the current generation of tools available to researchers simply cannot keep pace with the interactive, ad hoc nature of the data discovery process. So the discovery process is forced to match pace with the available software tools. Projects almost invariably progress slower than expected or never even get off the ground because the use of resources does not appear justified.

Data driven discovery in healthcare and life sciences is characterized by an interactive process in which researchers pose a question, analyze some data, and identify new correlations or insights that, in turn, raise new questions that may require new data, analyses and visualizations, which in turn often raise more questions to be explored. This researcher-led discovery process requires a software environment that not only supports but encourages the researchers' interactive journey as they follow the data where it leads them.

In this white paper, we will discuss how the scientific method can be supported by the discipline of data science to support iterative hypothesis, analysis and deduction to support discovery, what we call the Cognitive Hourglass.

Introduction

By now, organizations are well aware that data, sometimes big data and sometimes small, have the potential to revolutionize the effectiveness and economics of discovery in life sciences and healthcare. Ernst & Young[1] , Forbes[2] , and McKinsey[3] have all reported that better use of data has the potential to reduce drug R&D costs by $40 to $70 billion dollars annually, and the McKinsey Global Institute estimates that better use of data can create more than $300 billion in annual value across various life sciences disciplines.

Yet, to date, the actual benefits realized are falling far short of the potential. For example, although investments in pharmaceutical R&D continue to increase each year, the number of New Molecular Entities (NMEs) approved by the FDA has remained flat at only around 20 NMEs per year[4] . For every 5,000 to 10,000 screened compounds, only 250 will enter pre-clinical testing, about five will proceed to human clinical trials, and only one will gain FDA approval for marketing in the United States[5].

Despite more data, better technology, and larger investments in R&D, the discovery process has not yet delivered the expected results.

One reason is that despite the avalanche of data, it is still difficult for researchers to extract meaningful information from the data, information that leads to insight. The process of data-driven discovery often involves analyzing large and complex data sets, in different formats and with different schema. How do we know which data sets will be relevant and should be analyzed? How much work is required to include a data set in an analysis, typically requiring pre-processing, structuring, normalizing, analyzing, and visualizing the data? Which disparate data sets should be combined and compared to look for relationships and co-occurrences? Which methods should we use to analyze the data and visualize the results?

[1] http://www.ey.com/Publication/vwLUAssets/EY-beyond-borders-unlocking-value/$FILE/EY-beyond-borders-unlocking-value.pdf
[2] http://www.forbes.com/sites/emc/2014/03/17/how-big-data-is-transforming-drug-development/
[3] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[4] http://www.fda.gov/downloads/AboutFDA/Transparency/Basics/UCM247465.pdf
[5] http://www.phrma.org/sites/default/files/pdf/2014_PhRMA_PROFILE.pdf

The current generation of data analysis and visualization tools are designed for a process where the data sets, analysis types, and visualizations required are known a priori, in advance of the effort. These tools are built to support a linear workflow that looks like this:

Figure 1: Ideal data discovery workflow

But the reality of data science and discovery is that the data frequently takes us in new and unplanned directions. The initial analysis does not always yield the final result. Instead, it may provide some new insights that lead us to ask new questions that may require new data, new analyses, and new visualizations, which in turn may lead us to new hypotheses that raise new questions, and so on.

Figure 2: Actual data discovery workflow

We call this interactive process of data-driven discovery the "cognitive hourglass." It is the natural, researcher-led process of discovery, characterized by narrowing in on some result or insight that in turn leads us to ask new questions using new data sets to broaden along a different dimension that may provide another set of insights to narrow toward and so on.

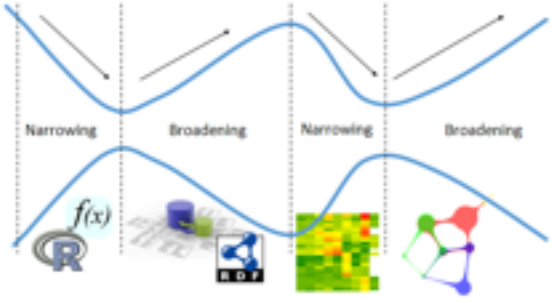

Figure 3: The Cognitive Hourglass

The Cognitive Hourglass – Two Examples

To make this idea less abstract, we present two real examples of how data-driven discovery raises unexpected questions and requires new data from different sources, new analyses, and new visualizations. Both examples involve cancer research. The first starts with gene expression data, traverses three data sets and multiple visualizations, and ends at journal literature from PubMed. The second starts with The Cancer Genome Atlas ("TCGA") data portal, leverages multiple open source technologies, and ends with some interesting statistically significant observations regarding particular gene mutation signatures. In neither case do we pretend that the research actually ends where our brief descriptions stop, and that is the point. Data-driven discovery and the cognitive hourglass are never that simple, and a researcher-led software platform must account for that.

The First Use-case: Leukemia and Gene Expression

In this example, a cancer researcher is looking for patterns in gene expression of leukemia in a patient study.

Initially, the researcher uses R[6] to analyze an RNA-seq[7] dataset containing levels of gene expression for 23,000 genes across the 23 different patients involved in a study. An interactive heat map, assisted by a word cloud of gene names sized by the variance of their expression levels, allows the researcher to see the variation and select from 23,000 genes the 65 that have the greatest variation in expression among the group.

[6] http://www.r-project.org/
[7] http://en.wikipedia.org/wiki/RNA-Seq

Figure 4: Narrowing of 23,000 genes to 65

Next, the researcher decides to query the NIH DAVID database for annotation information about clusters within the 65 genes that were gleaned from the heat map. The researcher is broadening his or her view along with new data in order to narrow again on the clusters.

Figure 5: Broadening out along a new dimension to include information from a new data source, the NIH DAVID database.

By including the NIH DAVID annotation data in the analysis, the researcher discovers that a cluster of 5 genes exhibit a cancer annotation (noted by the yellow highlighting and the corresponding Gene ID mapping in the visualization above). The researcher also discovers that the difference between the cancer variation and the normal variation is the difference of a single gene in the cluster.

Figure 6: Narrowing in a single gene expression responsible for the cancer variation

Next, the researcher can also see that two groups have significant published research associated with them, indicated by PubMed annotations contained in the DAVID database (shown as blue arrows linked to article identification numbers). So, after narrowing to certain gene clusters, the researcher chooses to broaden back out to the journal literature available from PubMed on one of those clusters.

Figure 7: Broadening out on the genes by incorporating all related PubMed research, then narrowing in on research associated with a cluster.

The occurrence of terms in the documents is displayed in a word cloud. Of interest to the researcher are the documents that mention leukemia and transplantation. Clicking on those terms in the word cloud incrementally filters down the list of articles until a single specific document is found.

Figure 8: Narrowing on just the research for two clusters that mention transplantation and leukemia

The researcher has traversed three data sets, proprietary and public, from different sources and in different formats. The researcher has processed them using a statistical package in R and then visualized results in a series of interactive visualizations as new questions arose.

The Second Use-Case: The Cancer Genome

TCGA is a massive, multi-faceted data portal for researchers to leverage clinical information, genomic characterization data, and high level sequence analysis of tumor genomes.[8] In this example, the cancer researcher is interested in exploring the co-occurrence of genetic mutations found in the tumor samples contained in the TCGA dataset.

Two genes that function in different pathways may both be found mutated in a given tumor sample, but genes that function in the same pathway are often found to have mutually exclusive mutation. If one gene is found mutated then the other is not, and vice versa.[9] Better understanding of gene mutation co-occurrence offers opportunities to better understand the pathways in which those genes play a role and the mechanisms by which different cancers override the healthy operation of the cell.

There are, however, so many different possible mutation combinations that in order to investigate co-occurrence, a researcher must first find a way to narrow the search space just in order to begin. So most data explorations begin with some form of feature selection, the process of selecting a subset of relevant features for use in construction of a model.[10]

In this case, the researcher starts by focusing in on a particular type of cancer found in the brain, Glioblastoma multiforme ("GBM"),[11] of which there are over 82,000 tumor samples in the TCGA dataset, containing over 847,000 distinct pairs of gene mutation co-occurrences. To further narrow in, the researcher then uses a technique called Fisher's exact test,[12] implemented in SciDB,[13] an open-source database optimized for such computations, to winnow the data set down to pairs of genes that are particularly statistically significant. These gene pairs are then plotted in an interactive chart that leverages the d3.js[14] javascript library to allow for in-browser interaction.

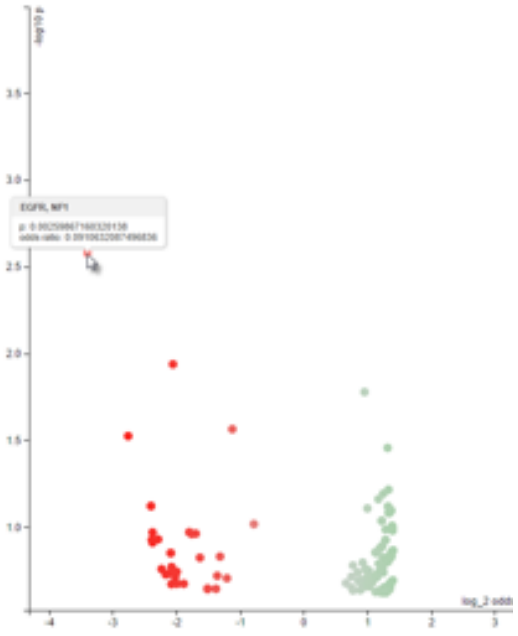

Figure 9: Plotting Fisher's exact test results on GBM gene pairs, filtered by top statistical significance.

[8] http://tcga-data.nci.nih.gov/tcga
[9] http://www.ncbi.nlm.nih.gov/pubmed/18434431
[10] http://en.wikipedia.org/wiki/Feature_selection
[11] http://en.wikipedia.org/wiki/Glioblastoma_multiforme
[12] http://en.wikipedia.org/wiki/Fisher%27s_exact_test
[13] http://www.scidb.org/
[14] http://d3js.org/

The red and green points shown in figure 9 represent the most statistically significant gene pairs, either because they were found to be both mutated in GBM samples (green), or because they were found to exhibit mutually exclusive mutation (red). The gene pair EGFR and NF1 stands out as the most significant pair. Because it is a mutually exclusive pair, there is some potential that EGFR and NF1 play a role in the same pathway and the pair is potentially useful for reverse-engineering how GBM infiltrates cells.

After narrowing in on this gene pair the researcher broadens out in a different direction, using the same Fisher's test analysis, restricting it to look just for the EGFR-NF1 pair, but expanding the search beyond GBM to look across all 1.7M tumor samples of all the different cancer types in the TCGA dataset. The results are again plotted.

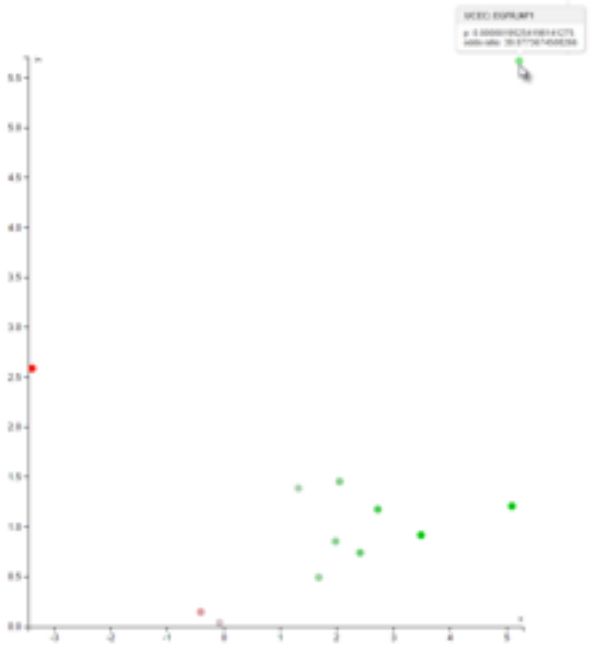

Figure 10: Volcano plot of EGFR, NF1 as they occur for any tumor, filtering for top statistical significance

In this case, the red and green points no longer represent gene pairs, but cancer types, and show the statistical significance of the EGFR-NF1 pair across different diseases. This broader perspective illuminates something interesting, that GBM is quite anomalous in this regard. It is the only type of cancer with a highly significant mutual exclusivity for the EGFR-NF1 gene pair. The only more statistically significant data point in figure 10 represents uterine cancer. But in that cancer the gene pair is found with co-occurring mutation instead of mutually exclusive mutation, suggesting that the pathways attacked by uterine cancer may be quite different than GBM, pathways in which EFGR and NF1 do not both perform critical functions in the same pathway.

Often the broadening and narrowing of the cognitive hourglass process involves moving along different dimensions of the data as the focus of the inquiry changes. So at this point the researcher returns from the broad view of all cancers shown in figure 10 to look again specifically at GBM, but this time from a slightly different angle. Now that EGFR and NF1 have been identified as genes of interest, the researcher uses a different analytical and visualization technique to make it easier to look at the particular genes involved in the pairwise mutations instead of just at the statistical significance of the pairs.

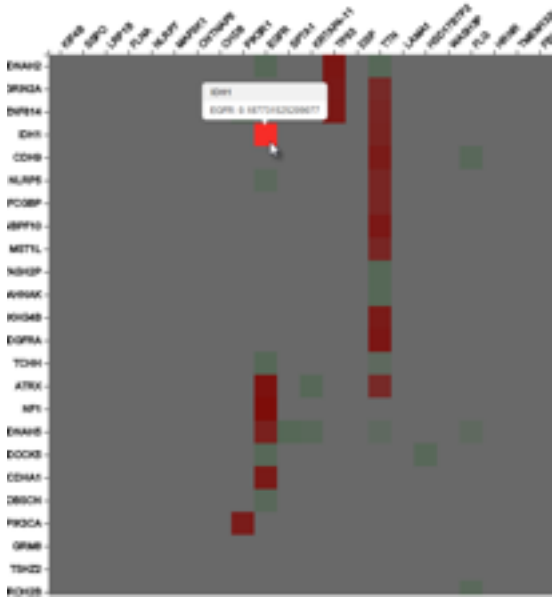

Figure 11: Identifying IDH1, EGFR, another mutually-exclusive gene pair involving EGFR.

By visualizing the same data contained in the plot of figure 9, but in a heatmap visualization that has been biclustered[15] using R, the researcher is able to quickly zoom in on the EGFR gene and all of the other statistically significant pairs in which it is involved. The heatmap shown in figure 11 was built using raphael.js,[16] another javascript library that makes it easy to support interactive visualization. The researcher clicks on the red cell representing EGFR-IDH1, another EGFR mutually exclusive pair. That click triggers another Fisher's test analysis in SciDB, broadening out again to look once more across all cancer types, but now in terms of EGFR-IDH1 co-occurrence.

The researcher returns to using the volcano plot in order to see how this new pair is represented across the other tumor types, and finds, unlike in the case of EGFR-NF1, that this time GBM is not the only tumor type showing strong mutually exclusivity for the gene pair. There is one cancer type with even stronger exclusivity for the pair, another classification of brain cancer called lower grade glioma ("LGG").

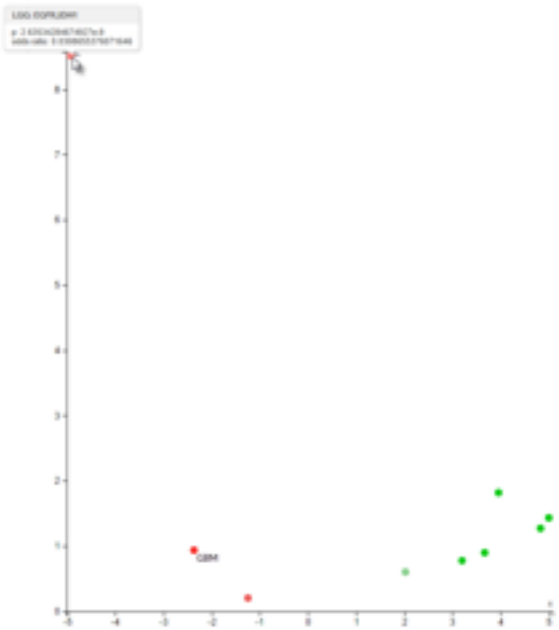

Figure 12: Volcano plot of EGFR, IDH1

The red point lower down the y axis in figure 12 represents a weaker exclusivity signal for the EGFR-IDH1 pair within a third tumor type: skin cutaneous melanoma. Now the researcher has a number of new angles that can be pursued, whether to explore the commonality between GBM and LGG tumors, investigating why the EGFR-IDH1 pair shows exclusivity in both diseases but the previous EGFR-NF1 pair did not, or to focus in on the melanoma samples to see whether there is really any commonality there.

Science is an iterative exploratory process and the goal behind this example was not to suggest any specific findings related to cancer but to show how dynamic the analysis process can be and how important it is for the tools that researchers rely on be able to keep up and move with them as their thoughts move in and out along different dimensions of the data, narrowing and broadening through the cognitive hourglass as they form, re-form, and evolve their hypotheses.

This Process can take Months or Days, Depending on the Approach

What makes the realization of these cognitive hourglass processes difficult in the current organizational and technological landscape is the significant and sometimes prohibitive friction between research needs and software capabilities. The reality is that most analysis tools and platforms today support a structured, software-led discovery process. The data to be analyzed often must first be pre-processed, formatted, and normalized. Prepackaged and custom made tools do what they were intended to do but multiple, fragmented tools are necessary to support different analysis techniques.

[15] http://en.wikipedia.org/wiki/Biclustering
[16] http://raphaeljs.com/

Figure 13: Workflow of the traditional, software-led discovery process[17]

Including new data sets in the analysis takes time, IT expertise, and resources. Different tools are required to perform different analyses. Since different tools are not integrated, it is difficult and time consuming to superimpose multiple data sets to identify relationships and correlations.

While traditional software-led analytics products are designed to model and answer a specific, bounded set of questions, they do not easily allow the researcher to follow a cognitive hourglass process, to easily navigate all the way to the ultimate result when the process is unpredictable. What is needed is a researcher-led platform where researchers can freely explore, iterate, and pivot across multiple data sets, analytical approaches, and data visualizations in a single unified environment, supporting the scientific and cognitive process as it unfolds

Exaptive: A New Approach Built to Support Researcher-Led Discovery

Exaptive is a data discovery platform that is built from the ground up to support and encourage the cognitive hourglass process, which in turn accelerates research and makes new discoveries possible where the barriers were simply too high with other technologies.

Key Features:

- Data interoperable, graph based architecture that encourages ad hoc inclusion of additional data sets from any source (e.g. database, web API, flat file), regardless of schema, by adapting without the need for further ETL or warehousing.
- Software inclusive such that researchers, data scientists, and software engineers can use their customary tools (e.g. R, Python, Hadoop, SPARQL, SciDB, SQL, d3, and other JavaScript frameworks) and lever age modules provided by Exaptive and Exaptive's user community, as the cognitive hourglass unfolds.
- Visualization extensible such that users can change perspectives and try unproven techniques with little sunk cost to see if new vantage points lead to a breakthrough.
- Reuse prior work by not creating schema-specific dependencies. Instead leverage data and tools modularly from one project to the next.
- Collaborate seamlessly (and securely) with team members, consortia, partners, crowd sourcing, or end-users. The researchers' cognitive hourglass is rendered in an easily deployable html5 webpage and is highly configurable for different roles and privileges.

Our platform is at work enabling the cognitive hourglass to progress and yield impactful insight in healthcare and life science R&D. We are:

- enabling a consortium of non-profit and for-profit organizations to collaborate to tackle brain disease with interoperable data;
- identifying potential false negatives in multiple sclerosis diagnostics in a matter of a few days where traditional data integration was taking months;

[17] http://www.hissjournal.com/content/2/1/3

- enabling a biobank to navigate through dozens of disparate research studies using its biosamples, showing it to be perhaps the single richest data set on multiple sclerosis in the world;
- empowering geneticists to identify new connections between genes and tumors in order to design breakthrough personalized;
- cancer diagnostics based on data an order of magnitude larger than the competition; and
- connecting clinical trial data to top-tier database hardware to enable researchers to explore adverse events and repurpose shelved therapeutics.

Conclusion

The technology exists to allow the ad hoc assembly of data, to analyze without being tied to a particular technology, and to experiment productively with new analysis tools and data visualization techniques. That is the way our minds want to work, the way data-driven discovery should work, and why a researcher-led discovery platform maximizes researcher productivity, shortening the discovery process, and enabling it to move forward when the barriers, technological or organizational, were too high before.

Two Ways of Working with Exaptive

Enterprises that want to leverage Xaps have two choices in working with Exaptive. The first is to license the Exaptive Studio for internal developers and data scientists to build the Xaps the organization needs. The second is to use Exaptive's internal experts to build them.

Contact us for a demo, free trial, or consultation.

About Exaptive

Founded in 2011 in Cambridge, MA, Exaptive offers a platform for creating knowledge from data that departs from traditional business intelligence and analytics software. "At Exaptive, we think less about giving people reports or dashboards, and more about giving them tools to explore a complex data landscape," says co-founder and Chief Executive Officer Dave King. By creating a software ecosystem in which cross-fertilization and code repurposing is a core feature, Exaptive is building a programming playground with the power to enable users' creativity and their potential for discovery.

Learn more at: http://www.exaptive.com