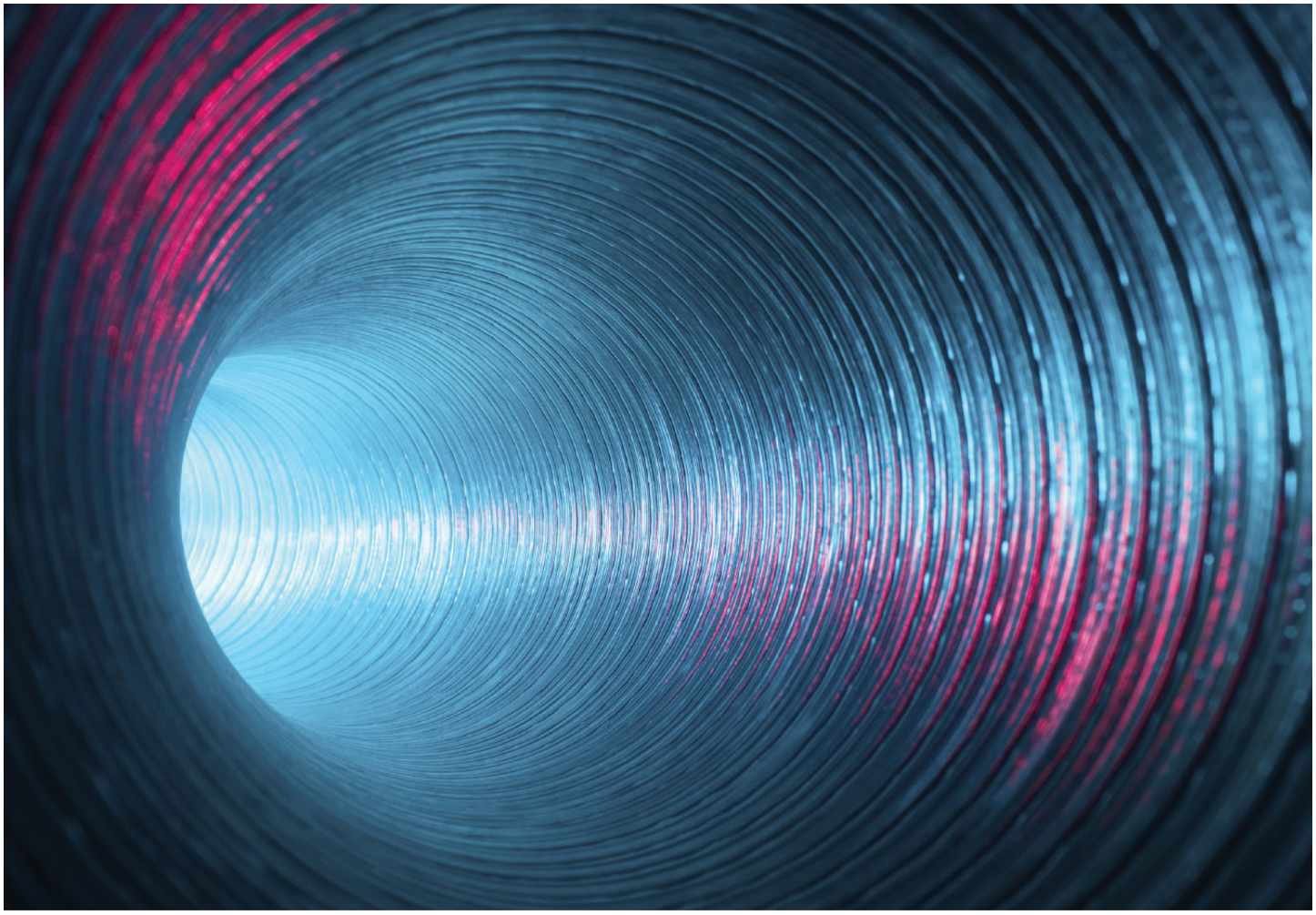


WHITE PAPER

BIG DATA AND THE NEEDS OF THE PHARMA INDUSTRY

JULY 2013

THOMSON REUTERS™

TABLE OF CONTENTS

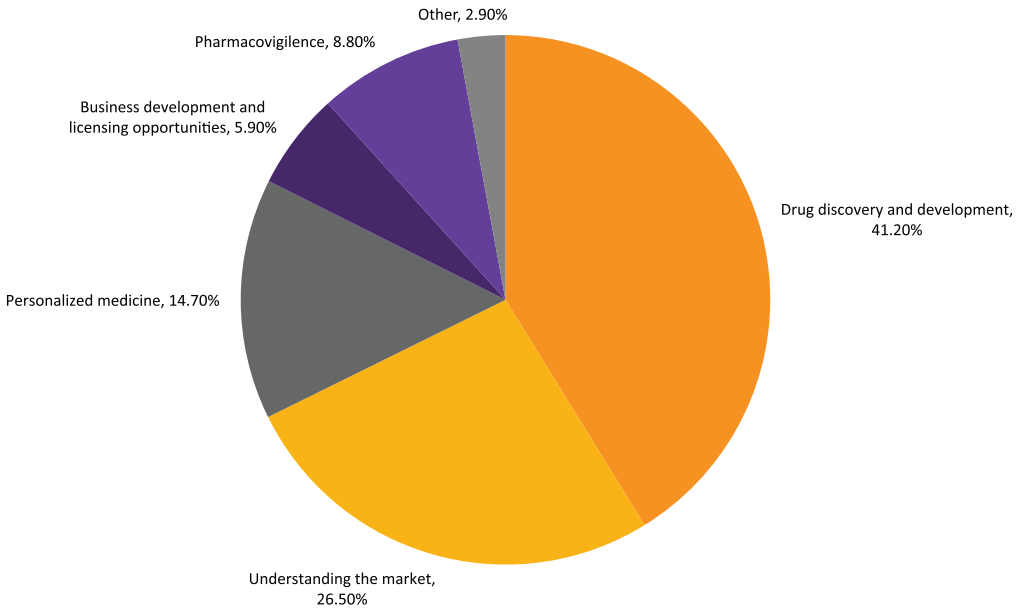
INTRODUCTION

Gartner, Inc. neatly defines Big Data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." Thomson Reuters Life Sciences sees Big Data as both a problem and an opportunity.

In a recent survey, Thomson Reuters asked a group of IT leaders in Pharma how they view Big Data. Unequivocally, 100 percent responded that Big Data is an opportunity. This isn't that surprising given that looking for empirical facts in large bodies of evidence has always been a driver of innovation. The Pharma industry generally works by identifying some correlation, e.g. between disease and protein, and then working out how to exploit it. The science as to why that correlation exists tends to follow despite recent trends towards science-led discovery. The ability to apply this thinking to even larger and richer sets of data is clearly an opportunity for Pharma to find new correlations and develop new drugs.

When asked about where they saw Big Data opportunities, our respondents overwhelmingly highlighted two areas of focus: early-stage discovery (41.2 percent) and understanding the market (26.5 percent). Figure 1 illustrates this, while also showing the emerging trend of personalized (or precision) medicine. Drug discovery has always been a data-driven activity and it makes sense to extend that to the new volumes, varieties and velocities of data coming out of the labs and out of public initiatives. Market understanding is new. It reflects both the change in the Pharma market dynamics, caused by the greater influence on prescribing behavior by payers, and the promise of rich patient-level data from electronic health records. Understanding the patient (personalized medicine), scoring 14.7 percent in our survey, is also a significant focus in Pharma right now, so access to, and the ability to, digest this kind of data represents a real win to our respondents in providing value to their businesses.

Figure 1: THE BIGGEST OPPORTUNITIES FOR BIG DATA



Source: Thomson Reuters Big Data Survey

BIG DATA INTO LITTLE DATA

The problem: humans can't work with Big Data directly. To realize the value of all this data, we need to reduce it to human proportions. When you examine what our customers are really doing with Big Data, what you see is the application of tools and techniques to "shrink" the data. In the Life Sciences business of Thomson Reuters, we call this "making Big Data look like Little Data." Little Data is the data we are equipped to handle. It comprises reliable, evidence-backed facts that scientists can use in models, visualizations and analyses. It is actionable data that can help Pharma companies with their core business of developing new and better drugs.

In drug discovery, this usually takes the form of designing in-silico experiments. This requires building up data sets from disparate sources, requiring cross-functional teams to work together on data that scores highly in the Variety vector of Big Data. Making this data act like Little Data is a challenge in data harmonization. You need common ontologies and ways to present data to scientists with very different skills and backgrounds, so it is delivered in a way they understand.

In patient understanding, the challenge is to get the data in one place and to filter the outputs so that the interesting correlations stand out from the "noise" of correlations that are either obvious or spurious--a situation analogous to signal amplification in the audio industry.

BIG DATA AS THE NEXT WAVE OF AN HISTORICAL TREND

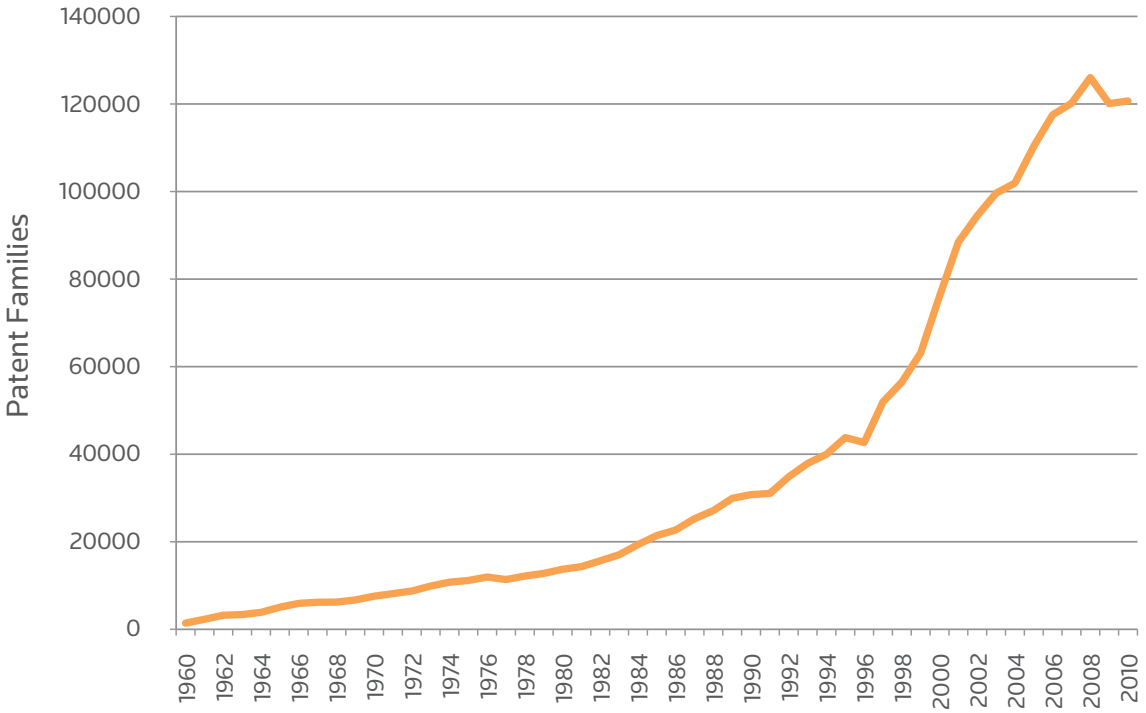
Challenging volumes of data is not a new phenomenon in Pharma. Rather, Big Data is an evolution in the application of data in drug R&D. It creates new challenges and provides new tools, but is not the complete revolution that some commentators claim it to be. Over many years, information companies like Thomson Reuters have supported customers in their efforts to divide whatever is the current "data elephant" into manageable chunks.

1960'S BIG DATA – PATENTS

An early Big Data challenge was the boom in patenting. In the first half of the twentieth century, a Pharma scientist could keep up with all the patents in his field by reading the patent applications him/herself. Figure 2 shows that by 1960, s/he would have to read over 1,000 patents a year to keep up. And, would have to be able to read at least English, German, French and Japanese. Abstracting and indexing services such as Derwent Publications emerged to address this challenge. Derwent's curation team read, classified and abstracted all the patents coming out of the leading patent offices around the world (as they continue to do so today). The editorially-enhanced information was published in weekly bulletins separated by area of interest so scientists could easily get to those of relevance.

As can be seen from the graph, this trend has continued more or less unchecked ever since. Nowadays nobody would even consider trying to keep up with patent information in any other way than setting up tailored alerting services on patent databases.

Figure 2: THE RISE IN GLOBAL PHARMA PATENTING



Source: Thomson Reuters Derwent World Patents Index

1970'S BIG DATA – CHEMISTRY

The 1970's saw the emergence of computer databases. In particular, the development of databases that could store, search and display chemical structures. These enabled Pharma companies to start to build the internal registry databases now counted as key assets and prime candidates for novel Big Data experiments. The emergence of online transactional services, the "Hosts," like Dialog, Questel Orbit (now called Orbit) and STN, enabled paper-based indexing services like ISI's Science Citation

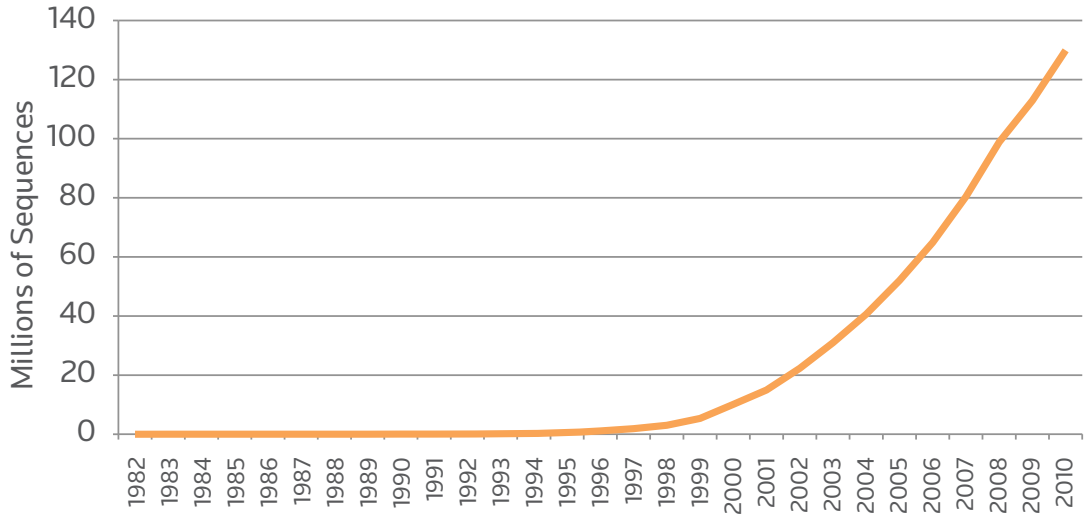
Index to become electronically accessible. This enabled the development of one of the first "data sciences"—bibliometrics. Having this data in a computable form enabled Pharma to apply a quality filter on the ever-increasing volume of journal articles.

Pharma companies were also eager to relate their internal registries to the broader pool of chemical information being generated in patents and journals. This was facilitated by the emergence of the key chemistry databases still vital to research today, including Chemical Abstracts Service, Beilstein (now Reaxys) and Current Chemical Reactions. Another new data science, "cheminformatics," was enabled by the availability of these resources. Cheminformatics made possible the discovery of structural motifs associated with efficacy or toxicity and subsequently enabled technologies like combinatorial chemistry.

1980'S BIG DATA – SEQUENCES

In the early 1980's, public sequence databanks such as GenBank, PIR and EMBL started to make available biological sequences identified by the nascent sequencing technologies. Figure 3 illustrates how the launch of the Human Genome Project kicked off an exponential growth in sequence publishing into these databanks.

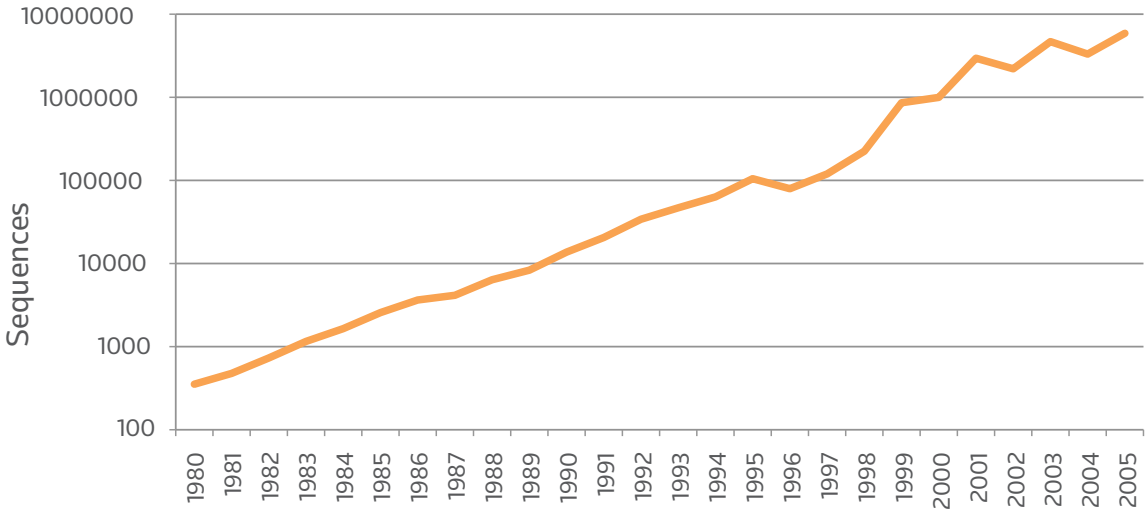
Figure 3: RISE OF PUBLISHED GENOMIC SEQUENCES



Source: GenBank

The growth in public sequence deposition was shadowed by similar growth in the patenting of sequence information, as shown in Figure 4.

Figure 4: 1980's BOOM IN SEQUENCE PATENTING



Source: Thomson Reuters GENESEQ

The sequence data explosion coincided with the emergence of personal computing, which gave rise to a new data science, "bioinformatics," developed in parallel with the new databases. Bioinformatics enabled this sequence data to be used for the discovery and study of new drug targets.

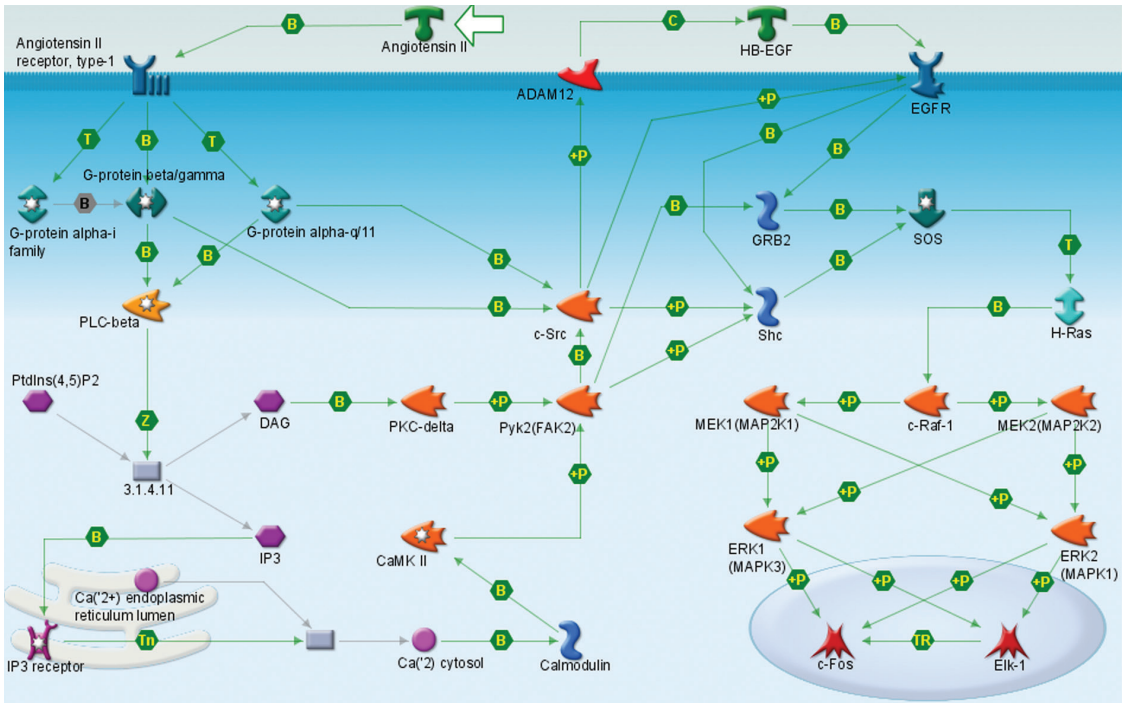
1990'S BIG DATA – ARRAYS

In the 1990's, microarray technology emerged as the new data challenge with the capability to simultaneously measure the expression levels of large numbers of genes, providing experimental data on an unprecedented scale. Using this technology, scientists could assess which genes were being turned on or off under different experimental conditions, e.g. samples treated with a drug vs. a placebo.

This generated extraordinary excitement within the scientific community. The reality, however, was that while microarray experiments could simultaneously measure expressions of tens of thousands of genes, experiments typically only comprised tens of samples. This resulted in a major statistical problem, with some gene expression profiles correlating with different samples purely by chance, and scientists got lost in the noise of Big Data. Bioinformaticians and biostatisticians were employed to help. The immediate solution was to turn to data reduction techniques like statistical clustering, typically employing techniques like principal component analysis and hierarchical clustering to reduce complexity. However, these statistical approaches didn't capture "why" the gene expression changes were important, from either a biological or pharmacological perspective.

In 2000, like Figure 5 shows, GeneGo's Metabase platform was developed to capture the relationships between genes as biological pathways and maps. Using teams of human curators extracting and relating high quality information from the scientific literature. This pathway approach soon became known as "systems biology." The GeneGo team added further invaluable information to their pathways, including where in the body the genes were expressed, associated diseases, drug targets and biomarkers. Scientists could now take the immensely complicated output generated through microarrays, and use it to better understand disease biology, identify new drug targets and biomarkers.

Figure 5: PATHWAY MAP OF ANGIOTENSIN'S ACTIVATION OF ERK VIA TRANSACTIVATION OF EGFR



Source: Thomson Reuters MetaCore

2000'S BIG DATA – NEXT GENERATION SEQUENCING

The ability to sequence a genome, which drove the 1980's Big Data challenge, is now creating a new one. New technologies have increased the speed of sequencing and driven down costs. What took the Human Genome Project 10 years to do for one (consensus) genome can now be done for an individual in days. Routine sequencing of patient genomes in clinical trials is just around the corner. Systems to manage the raw data are maturing, but the tools to make sense of this data are still in development. These tools will make possible the idea of personalized medicine. But, scientists and researchers need reliable information about variant-disease and variant-response information. This is a classic example of where curated Little Data can be used to make sense of Big Data coming out of the next generation sequencing (NGS) machinery.

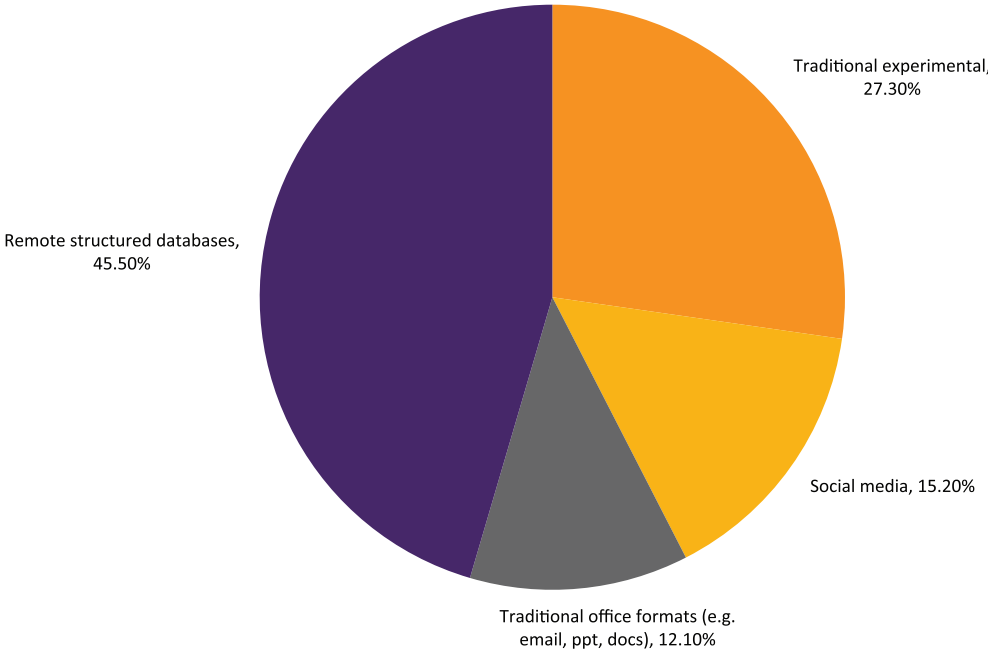
The new Gene Variants API from Thomson Reuters makes it possible to process chromosomal location information and get a summary of the kinds of disease susceptibilities and/or therapy response profiles associated with the identified variants. GenoSpace is using this to create patient summaries that are used by physicians to make more informed decisions.

BIG DATA IN THE PRESENT

As we have seen, each decade has presented a new data challenge, and information providers have played a key role in turning the Big Data into Little Data, so data scientists can turn the data into value for their companies. Today's Big Data challenge is one of diversity. This time, it is not unmanageable volumes of a single data type that concern Pharmas, but rather it is how to connect data from multiple places, internal and external, to feed the innovation process.

As seen in Figure 6, respondents to our survey confirm this. Their biggest challenges are integrating external database content (45.5 percent), internal experimental content (27.3 percent), external social media content (15.2 percent) and internal document content (12 percent).

Figure 6: THE TYPE OF DATA PROVIDING THE BIGGEST CHALLENGES



Source: Thomson Reuters Big Data Survey

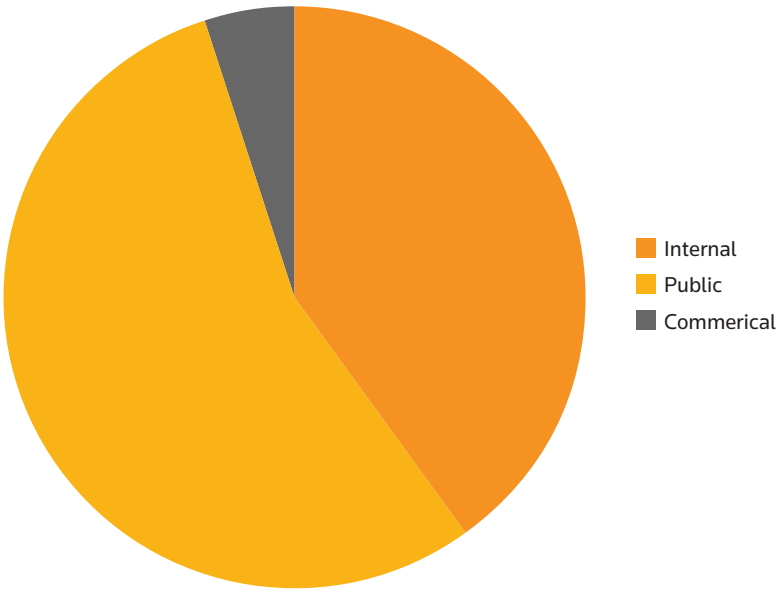
NEW TOOLS AND TECHNIQUES FOR OLD PROBLEMS

One hope for Big Data technologies is that they will assist in combining data in a way to make it actionable. Our survey showed a strong interest in NoSQL databases, linked data/semantic web technologies and visual analytics as tools they expected to use to solve these challenges.

Insight from internal data

The volume and variety of data behind the firewall of a typical Pharma company is fast approaching a point where it exceeds the content from the outside, as demonstrated in Figure 7.

Figure 7: CONTENT USED IN PHARMA R&D



Source: Thomson Reuters

Many organizations today have a hybrid data model, where they purchase the quality content externally as well as generate a lot of data internally. All of this data (much of it textual and difficult to read) needs to work together, whether from inside or outside the firewall. And they need ways to unlock the value in this data as well as the ability to do it in an agile fashion to construct a system in days rather than months that answers a specific question and can be altered on the fly.

Component-based systems like Accelrys' Enterprise Platform, Pipeline Pilot, have become the "Swiss Army Knife" of informaticians to do exactly that. As illustrated in Figure 8, it brings external data in, acting as the bridge between these two worlds. Accelrys partners with Thomson Reuters to provide a library of Pipeline Pilot components that offer convenient plug-and-play access to the Thomson Reuters Cortellis for Informatics Web Services APIs.

Thomas Mayo, partner and developer evangelist, Accelrys, is acutely familiar with this plug-and-play scenario. His clients are the chemists, researchers, developers and informaticians processing life sciences data. They are pulling in content from multiple sources—internal and external—and analyzing it to improve decision making. As he says, "The value our clients get is in the ability to cross-reference data and ask more questions of it."

A tool on top of Accelrys' Pipeline Pilot enables users to scan billions of data points from a multitude of sources, thereby informing the drug discovery and development process. It enables companies to unlock the value in the data being analyzed.

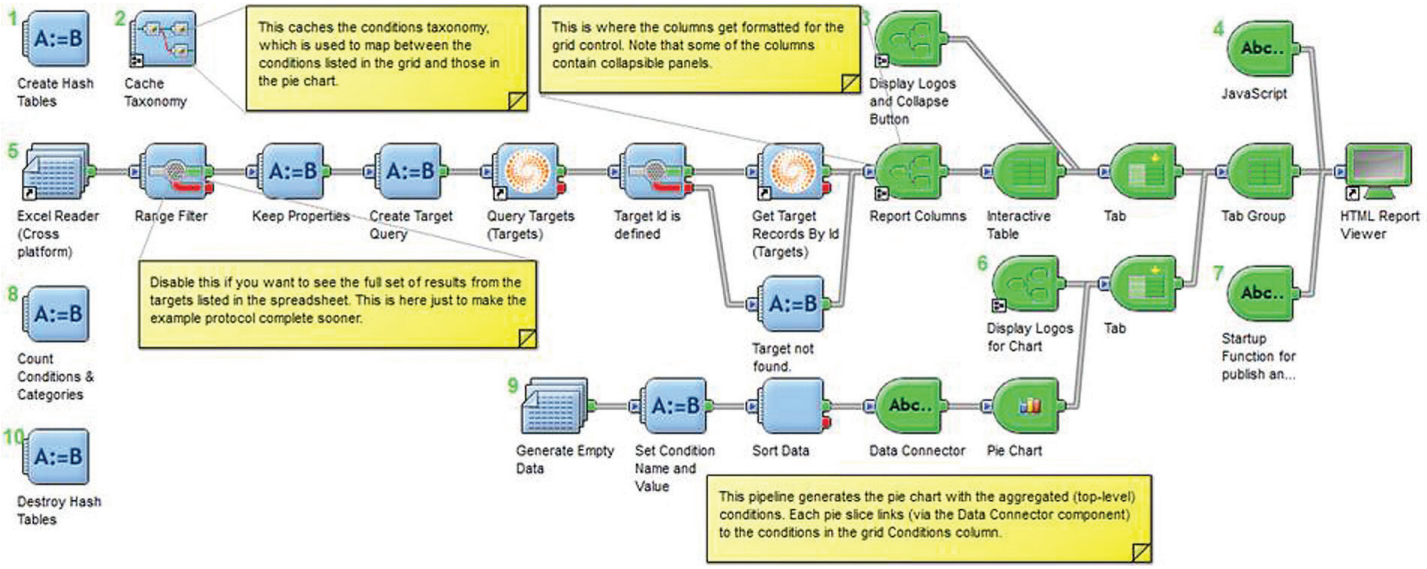
"Thomson Reuters ontologies and APIs give our users the ability to cut through the massive volume of data and get meaningful insights, answering questions around failure rates, executing high impact exercises, identifying the compounds worth keeping, and so on," says Mayo.

There are numerous data points that are important in life sciences along the value chain. It is important to know who is invested in what, how the pathways work, where the IP is, what statistical analysis has been done, how genetics come into play, and the like. Knowledge is power and information fuels knowledge.

Accelrys is not just working in the discovery space. Pipeline Pilot is the only scientific informatics tool being utilized in the National Health Service in the UK. Now that they've built components to query that data (57,000 patient records), they can query for outcomes needed by those developing drugs by age, population, and so on.

Looking ahead, Mayo sees a lot of possibilities in the combination of Accelrys' tools with curated databases. "We can build better applications with services and software, like competitive intelligence dashboards, to know what is in the pipeline, where development funds are going or what drugs are performing well in the market. Once people get a handle on the data Thomson Reuters has and the type of questions they can ask, they point to where they see they need more of it. We partner with Thomson Reuters because it is the world leader in content curation."

Figure 8: EXAMPLE INTEGRATION OF TARGET DATA FROM THOMSON REUTERS INCLUDED WITHIN ACCELRYS PIPELINE PILOT



Source: Accelrys/Thomson Reuters

Insight from external data

Over 45 percent of respondents to our survey reported access to external content as their top challenge. The complication there is that each data provider, public or commercial, presents his/her content in a different way, via their own Web site, using terminology and user-access-control mechanisms unique to their organization.

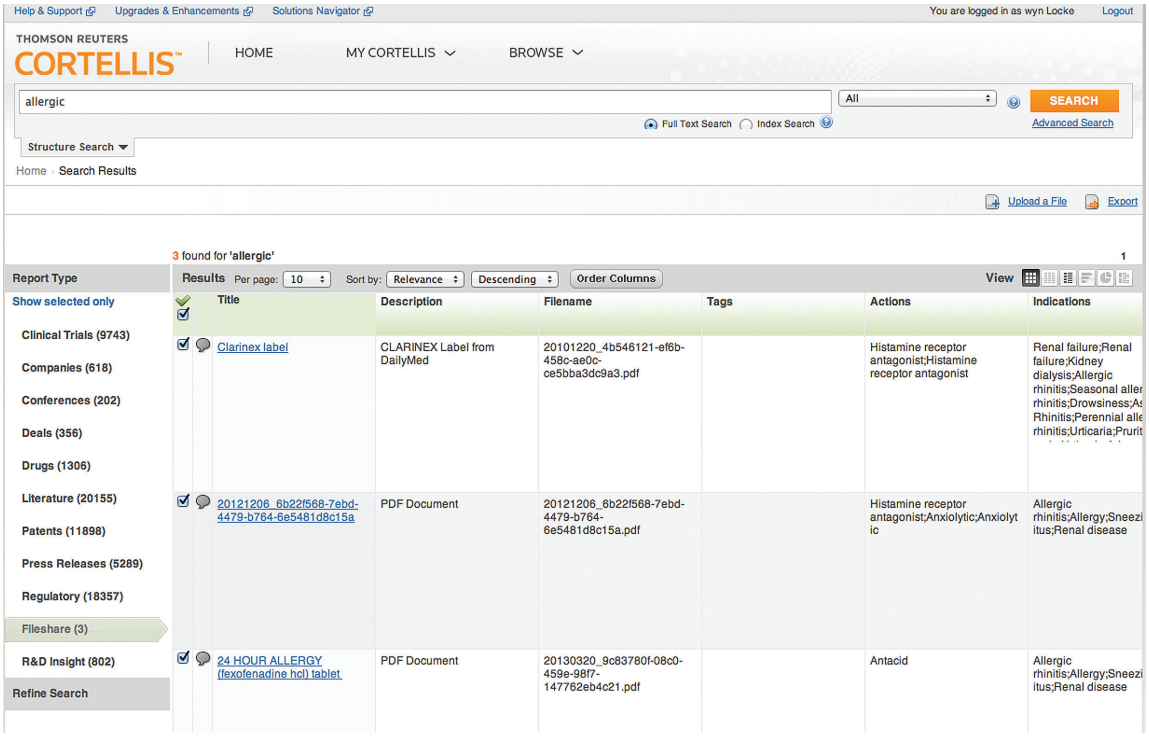
Information professionals have become adept at selecting the most valuable resources, extracting the most relevant content, and then reformulating and presenting the results to their users. But, this process can create an organizational bottleneck that isn't sustainable. Scientists who would most benefit from this are put off by the time it takes to do the analysis and the inconvenience of having to explain their requirements. The result is often that the user performs a naïve search on a Web site, finds nothing and feels satisfied that he has freedom to operate – only to discover months later that this is not the case.

One way around this is to deliver high quality information to users by "mashing up" different APIs, e.g. using Accelrys' Pipeline Pilot. Figure 9 shows another approach—taking advantage of the Web portal interface of Thomson Reuters Cortellis.

Cortellis is designed to make available a collection of content from previously independent Web products in a single end-user-friendly interface. It is a natural extension to look at integrating non-Thomson Reuters content in the same interface. This first came into play a few years ago with the launch of the Pipeline Data Integrator. The Pipeline Data Integrator combined and mapped the content from the leading pipeline databases so users could get a consensus view from all these resources in a single search.

Cortellis for Information Integration takes this a step further. Customers can look across commercial databases and also load public databases and even their own content. A single search shows results for all the content sets and leverages Thomson Reuters ontologies to make the content findable. The end-user scientist no longer has to miss that vital resource. Content on Cortellis for Information Integration, wherever it comes from, is also able to be annotated so users can capture the organizational "secret sauce" (internal knowledge) right alongside the rest of the content.

Figure 9: CORTELLIS FOR INFORMATION INTEGRATION SAMPLE SEARCH ACROSS PROPRIETARY AND THOMSON REUTERS CONTENT



Source: Thomson Reuters

Insight by combining heterogenous data sources

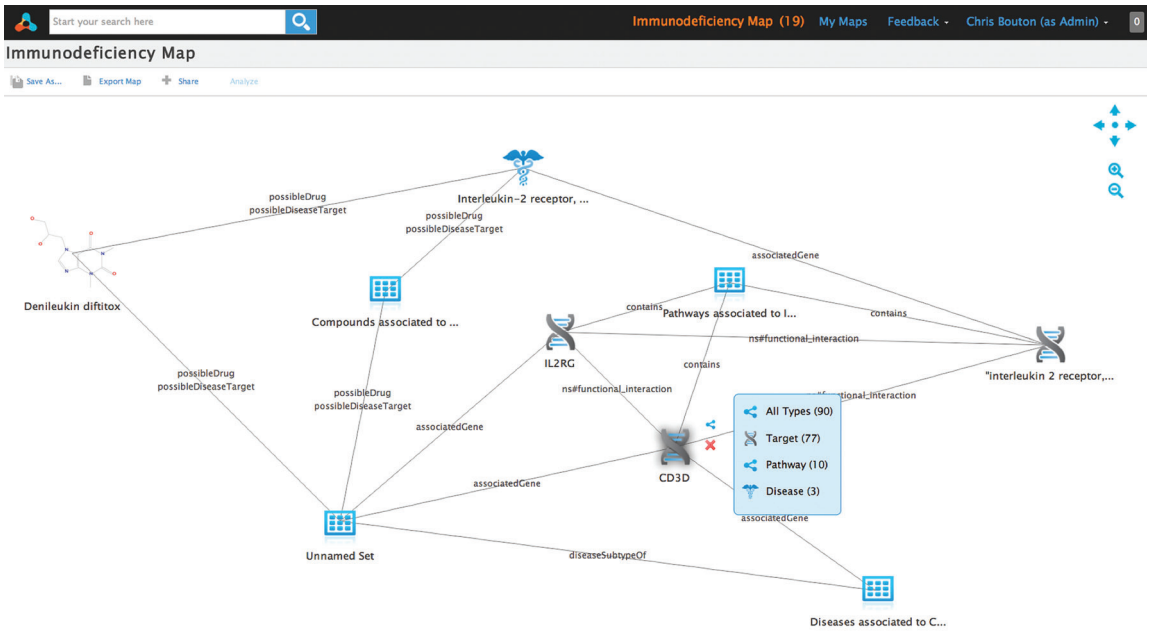
Linked data provides a powerful framework for integrating internal and external content in a way that enables users to ask questions of the combined data in a single "master graph." Entagen is doing this, powered by the Cortellis for Informatics APIs, illustrated in Figure 10. Entagen's EXTERA (semantic data) technology is used to integrate and interact with data from across the organization, both structured and unstructured content.

Chris Bouton of Entagen says it's all about "connecting the dots in Big Data." "In the life sciences, it isn't just the understanding of individual drug entities, pathways, etc., it is the associations between them that are really important to scientists to generate the right hypothesis, advance the right projects, and so on."

Customers are using EXTERA in precision medicine, research (early discovery through screening and development), competitive intelligence, and in the legal sector where a lot comes down to understanding the entities, people, and case documents. Entagen's KNOWLEDGE MAPS provide powerful ways of interacting with the information.

Asked about the value of curated Little Data in Big Data, Bouton says: "It's critical. The reason is that there is a classic pyramid from data through information to knowledge. Everyone talks about Big Data. We certainly have more data than we know what to do with. There is a need to get the data into a framework so that you have information. Adding patterns to give data meaning. Without high quality data like that which is provided by Thomson Reuters, you're lost. Next, you need to be able to translate that information to knowledge. Without any background context which one can rely on, this process is severely hampered, if not impossible."

Figure 10: IMMUNODEFICIENCY MAP (INCLUDING THOMSON REUTERS DATA)



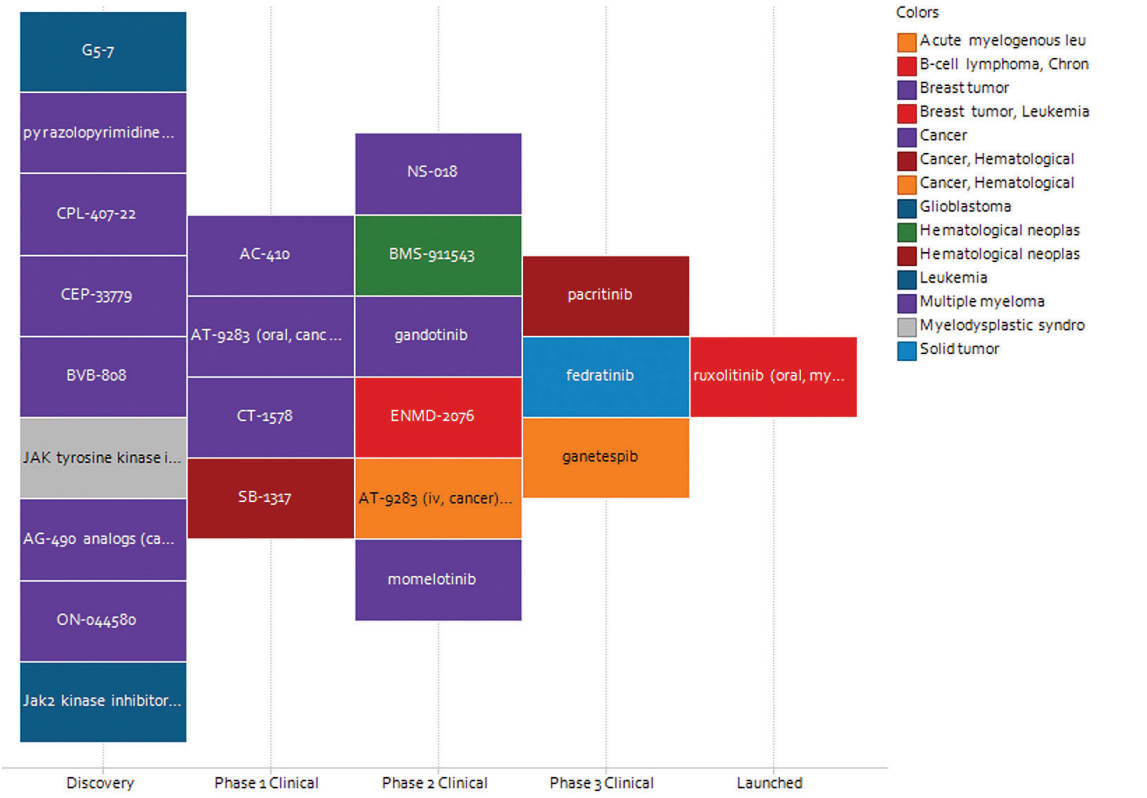
Insight through interactive visualizations

One of the best ways to get value out of Big Data is to be able to visualize it, but merely putting all the data into a nice chart is not always enough. The new generation of visual analytics let the user interact with the data, choosing what range(s) is plotted, filtering out items that are perceived to be relevant and deriving a tailored view that can then be used in presentations and reports. These tools also allow skilled information designers to template analyses, like Figure 11, that simplify the tool for non-specialists while providing the same analytics capabilities.

In the course of a collaboration between Thomson Reuters and the Oncology iMed of AstraZeneca, Thomson Reuters designed a set of templates using TIBCO's Spotfire software focused on a set of common questions that arise in the clinical stages of drug development.

- What does the competitive pipeline look like in this indication or for this target?
- How do the safety profiles of these drugs compare across the clinical trials they have been tested in?
- What is the likely duration of my trial in comparison with similar trials?
- Is a competing trial likely to complete before mine?
- How has this drug been progressed through clinical trials?
- Which indications are the drug owners prioritizing?

Figure 11: THE PORTFOLIO VIEWER WITH VISUAL DISPLAY OF THE COMPETITIVE LANDSCAPE



Source: Thomson Reuters Decision Support

NEW PROBLEMS; NEW OPPORTUNITIES

Social media

Social media is a new dimension to Pharma but one that 18 percent of respondents said they work with daily and 15.2 percent said they think of as a significant challenge. Jochen Leidner, lead research analyst from Thomson Reuters, specializes in studying social media. "It's true that people tweet about medical worries, symptoms, the drugs they take and their side effects," he says. "Social media has become such an important factor. Even Google famously used Big Data analytics to out-perform the Center for Disease Control in predicting the spread of winter flu in the United States."

Pharma companies are looking at social media to meet their goals of better understanding the patient. They can use social media to find out about unmet needs, new adverse events and patient compliance issues. This information isn't just prevalent in the billions of daily Tweets and posts in social media in the West; there is also a huge surge in the specific social media services in emerging markets.

Thomson Reuters partners with a social media aggregator, DataSift, to provide social media analysis for its customers. DataSift makes it possible to query across the various sources while Thomson Reuters applies the domain knowledge and the ontologies to draw insights from the data. "Think of it like a water purifier," says Leidner. "If you are in the desert, you need to consume water and need it to be clear and clean so you don't get a disease from it. Thomson Reuters is like the purifier of social media data. We filter out that which is of value and that which they care about. No one else can do it in such a specialized way."

Personalized medicine

Since we are at the point of being able to sequence a human genome for less than $1,000 USD over a couple of days, physicians look to leverage the insights they can glean from genomic analysis to provide better patient diagnoses and care.

Yet, as the amount of genomic data increases, so too do the challenges around storing, managing and securing all this information. Luckily this dilemma has a solution. It lies in the utilization of data-management resources that parse the unfathomable amount of information available to find the

nuggets of information that are most relevant.

John Quackenbush, CEO of Genospace, an information architecture company focused on genomic medicine, is no stranger to the challenges of gathering, storing and managing Big Data. "When I look at someone's genome today, I find thousands of variants in coding sequences. Regardless of what I'm trying to determine, it is hugely important to put all this data into context. Thomson Reuters has exquisite content for studying gene variants and pathways, looking at mutations and the broader landscape in the progression of disease. The Gene Variant Database is an incredibly useful resource for providing accessible gene diagnostic information. It allows me to remove the noise and pull out the signal."

All of this gets to the evolution of precision medicine, which is using the information and technology available today to manage each patient's healthcare according to their genetic background. As Quackenbush says, "The metadata is really important. As we look at a disease like cancer, a mutation shows options for what might be a better treatment for the patient. It is the information in context. Thomson Reuters is really good at providing context specific information. Not all mutations are equal."

The Thomson Reuters data sets enable scientists, researchers and doctors to put the most valuable, essential information into context. GenoSpace works with a number of its clients to bring the insights from genomic data to life. One such example is a pathology lab looking for copy number variations related to diagnostics in cancer. GenoSpace synthesized reports and genome type information into a database, pulling information from the Gene Variant Database to give the lab additional information on mutations and other aspects to enable the researchers to draw better conclusions. The data, in its proper context, allows physicians to more accurately diagnose and drive interpretations.

"By investing today in the highly curated data from Thomson Reuters, we are insuring ourselves from the aftermath of the tsunami of data coming in the future," says Quackenbush. "Having a trusted source to reduce complexity is extraordinarily useful. The amount of information in the future will continue exploding. We need to catch the wave now."

For big pharma, there are challenges around understanding the competitive landscape and running more targeted clinical trials. The pharmaceutical companies need to know what is already in trial, at what stage, and the success/failure rates. "Keeping up with all of this as the number of mutations increases is a monumental challenge. The Gene Variant Database and Integrity, both from Thomson Reuters, are really the elite solutions in these areas," says Quackenbush.

It won't be much longer before everyone in clinical trials is routinely sequenced, trumping the prior genetic markers such as height, weight and family history for diagnosis and treatment. The promise with NGS is that you know a lot more about the people in the trial and can start to see patterns and look at gene mutations that could cause ill side effects.

Together, Thomson Reuters and GenoSpace are leading the way in transforming the Big Data challenges of next generation sequencing into actionable insights for physicians making treatment decisions.

DATA ABOUT BIG DATA

Knowing the questions and how to answer them

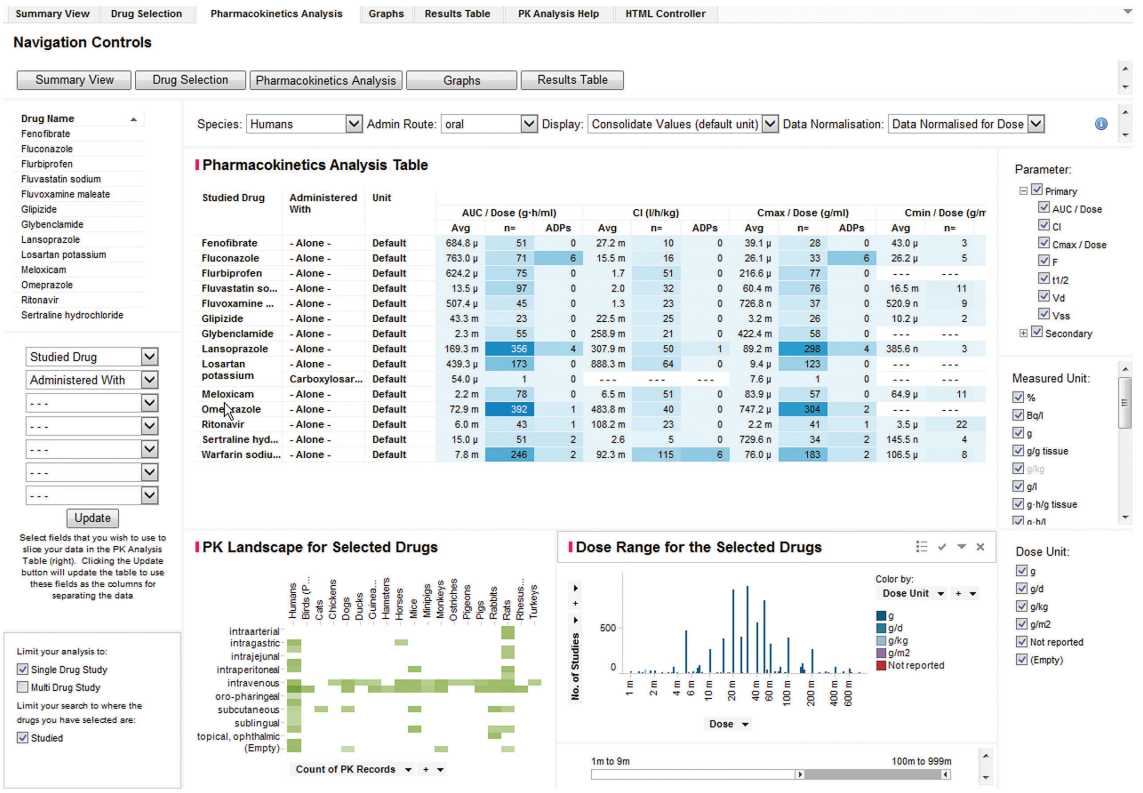
We believe that the next area of need after Big Data is going to be Big Questions. There is a huge amount of knowledge encapsulated in the process of asking the right questions and getting the right answers to those questions. There is a synthesis here between the understanding of what the "real" question is, the understanding of where the data can be found (and the strengths and weaknesses of each source) and how to analyze the data to make it digestible by the eventual user. These skills have normally been the role of information professionals but there is now a shortage of them and a time lag introduced by the process. Through the Thomson Reuters Life Sciences Professional Services team, there is a deeper understanding of the common questions its customers ask. Emerging from that understanding, Thomson Reuters is finding new ways to encapsulate the content and analytics in a way that goes directly to the underlying need.

In some cases, there is still the need for an intermediary to work with the end-user, and with the data and analytical tools. This is especially true where some combination of internal and external data is required to answer the question, such as in drug repurposing exercises. Nevertheless, Thomson Reuters has found that there are reusable components in the information processing, and have used that to turn around these requests in a timely manner. Specifically for the task of drug repurposing, Thomson Reuters built an internal toolset that pulls together all the relevant information in a single interactive "workbench" that its experts can work with.

In other cases, Thomson Reuters is developing tools to be used directly by the end user. These new capabilities on the Cortellis platform (Figure 12) are generating solutions to support research and development. From drug metabolism and pharmacokinetics scientists looking for solutions to absorption, distribution, metabolism, and excretion problems, to business development and licensing professionals looking for non-obvious partners in licensing drug candidates.

Aggregating pk data for a selected group of drugs, Figure 12 shows how dmpk scientists can immediately review the typical pk performance, normalized for dose—and review the change in pk or resultant drug-drug interactions when co-administered with other drugs.

Figure 12: THOMSON REUTERS CORTELLIS PLATFORM AGGREGATION



Source: Thomson Reuters

BIG DATA ENGINEERS

Big Data -> Little Data -> Insight requires specialist skills

In the world of Big Data, it is easy to become seduced by the technology and not address the human issues. Converting Big Data into Little Data and into actionable content is a skill. The skill set of a Big Data engineer includes understanding how to acquire, order, map and visualize content—plus the scientific knowledge to know what kinds of activities are useful in support of the questions pharmaceutical companies will ask.

Big Data opens up a world of new opportunities for data mining and reasoning. Through a partnership with The Imperial College of Science, Technology and Medicine in London, UK, Thomson Reuters is exploring how to best add value in a changing content landscape through engineering and application of new Big Data technologies. These technologies can be grouped into three areas:

1. Infrastructure and thinking on the petabyte scale
2. Knowledge representation, with the shift away from relational structured (SQL) to unstructured (noSQL), including new approaches like RDF and linked data
3. Deep learning and reasoning, with machine learning approaches less prone to over-fitting

Yike Guo, professor in computing science at Imperial, makes the point that "Big Data is all about VALUE not SIZE. You can have a petabyte of rubbish; it's the way that we treat it that is crucial." Content providers recognized, way before anyone else, that data could be a resource like oil that we can produce and refine to add more knowledge. However unlike oil, it's a resource that will never be used up. It's endless and you can keep adding to it. Thomson Reuters Life Sciences content is like the highest quality oil. What's new is the concept of the data 'product' (not provider) industry. It's not just data anymore."

The collaboration between Imperial and Thomson Reuters covers:

1. **The content itself.** With different levels of refinement, assimilate the content. For example, compare the low level molecule data with the high level of a pathway system.
2. **Building a simple data resource pool.** Working together, carry out data-exhaustive research. Create a universal search engine building connections between genes, diseases and other areas, creating a complete search experience.
3. **The changing scientific delivery mechanisms.** Thomson Reuters and Imperial present knowledge in the most succinct way and reference the provenance.

In training the new generation of data engineers, Guo says, "When we're thinking about data engineering, we're thinking about what the data landscape will be in five years time. So the issues we work towards solving now are not necessarily today's challenges, but they are real problems that will exist in the future. By trying new technologies and solutions now, the industry will be able to commoditize them in the years to come."

This can be looked at in three ways:

1. **Storing data.** "This is a challenge around infrastructure. We need to be thinking about petabytes as a starting point and basing our designs on petabytes. For example, the National Phenotype Centre carries out metabolite profiling in humans. They process 20,000 samples/year, (3.2 petabytes per year). And we're thinking about assay, omics data from high-throughput screening –mixed with known sequence type stuff."
2. **Knowledge Representation.** "Presenting pathways, that's extremely interesting. The model is not relational, we need models like RDF, linked data and pathways, technologies like Neo4J."
3. **Deep Learning/Analysis.** "Machines and algorithms that allow for new experiments. By understanding what a question means, we can structure the content to answer it. We need a new generation of machine learning algorithms, using tools like neural networks. Over-fitting is suddenly not such an issue with Big Data and as we have more computational power."

CONCLUSION

Going back to Gartner, it is now official that Big Data has already crested the "Peak of Inflated Expectations" and is going down the "Trough of Disillusionment." For most IT groups in Pharma, this is a relief. Pharma has had more than its fair share of inflated expectations and needs to get on with the job at hand. That job is mostly about getting the right data to the right people in a way that they can use it to do their jobs smarter. The information providers need to recognize that the Big Data opportunity is not about the technology per se, it's about learning to work in an environment where the data (Big or otherwise) has to flow to be relevant. That means breaking down the walls between internal, external, public and commercial content and giving customers the tools to work across them, whether that is by opening programmatic access, acting as a data warehouse, providing professional services or designing intuitive analytics.

Thomson Reuters believes that it isn't any one of these approaches that is the "magic" solution. Each customer problem has its own dimensions and will require its own solution. The key is to build the components that turn that from a bespoke development solution to a tailor-made, off-the-shelf one.

What enables this is Little Data. Little Data has the ontologies that can glue together content from multiple sources. Little Data can provide the noise filter that makes sense of Big Data analyses. Little Data can provide the context that enables users to follow a train of thought through the data, and Little Data can power the analytics that turn content into insight.

Information providers have a history of building and maintaining this Little Data. It's a great opportunity for Thomson Reuters and its customers and partners to build the next generation of solutions that will enable the discovery of new and better drugs which, after all, is what this is all about.

ABOUT THOMSON REUTERS

Thomson Reuters is the leading source of intelligent information for professionals around the world. Our customers are knowledge workers in key sectors of the global economy. We supply them with the intelligent information they need to succeed in fields that are vital to developed and emerging economies such as law, financial services, tax and accounting, intellectual property, science and media.

Our knowledge and information is essential for drug companies to discover new drugs and get them to market faster, for researchers to find relevant papers and know what's newly published in their subject, and for businesses to optimize their intellectual property and find competitive intelligence.

NOTE TO PRESS:

To request further information or permission to reproduce content from this report, please contact:

Laura Gaze
Phone: +1 203 868 3340
Email: laura.gaze@thomsonreuters.com

For more information, please visit
go.thomsonreuters.com/big_data or email **cortellis@thomsonreuters.com**

THOMSON REUTERS REGIONAL OFFICES

North America
Philadelphia +1 800 336 4474
+1 215 386 0100

Latin America
Brazil +55 11 8370 9845
Other countries +1 215 823 5674

Europe, Middle East, and Africa
London +44 20 7433 4000

Asia Pacific
Singapore +65 6775 5088
Tokyo +81 3 5218 6500

For a complete office list visit:
ip-science.thomsonreuters.com/contact

LS-201307

THOMSON REUTERS™