The life sciences have always produced large amounts of data. As the amount of data produced grows exponentially, new methodologies are needed to glean useful insight from it. Cloud computing company Code-N was founded to solve Big Data problems that can't be solved with the keyword-based Internet software and search technologies. Code-N's technology spans the gamut of life science applications helping pharmaceutical and biotechnology companies more quickly and more efficiently discover new drugs, repurpose old ones and monitor the current slate of already commercialized products for both safety and competitive reasons.

Code-N founder and Executive Chairman Marketta Silvera has more than 20 years experience as chief executive of four technology companies serving the health care and financial industries. Her Big Data background includes work with life science inventors and cheminformatics experts at the Netherlands Bioinformatics Center several years ago to develop concept-based technology solutions that leverage public/private partnering and Big Data.

Code-N is led by CEO Randy Haldeman. Before becoming Code-N's CEO, Haldeman led the content division of Symyx/Accelrys (NASDAQ: ACCL), a leading provider of informatics solutions to more than 1,300 corporations in the pharmaceutical and biotechnology industries including companies such as Merck, Bristol Myers Squibb, Pfizer, Eli Lilly, Novartis, AstraZeneca, and GlaxoSmithKline.

Marketta and Randy took some time recently to share their thoughts about Big Data with Carlyle & Conlan's Don Alexander, for their 2014 Report.

Don: ***What is your definition of big data?***

Marketta: *Big Data is for real and it's exploding in our digital lives. Since the late 1990s, assisted by the Internet, the world's businesses and population have freely participated in generating new data in numerous forms. By now there are many definitions for Big Data. The consulting firm NewVantage Partners' "Big Data Executive Survey 2013" defines Big Data as "collections of data so large, complex, or requiring such rapid processing (sometimes called the volume/variety/velocity problem), that they become difficult or impossible to work with using standard database management or analytical solutions."*

But it's the volume of Big Data that's unimaginable. Consider that just a year ago, the world was creating some 2.5 exabytes (25 billion gigabytes) of data every day, which on an annual basis is equivalent to filling 30,000 new U.S. Libraries of Congress. TechAmerica Association estimates that 90 percent of the data that has ever existed has been created in the past two years. Internet searches, satellites, massive research projects like the human genome, mobile devices, security cameras and remote sensors – all these data generators and dozens more, are fueling data proliferation on an epic scale.

Big Data has become the new raw material in business, next to capital and labor. Today's big market opportunity lies in innovative technologies that help extract the intelligence and relevant insights from overwhelming quantities of information. The benefits of accessing this vast resource are extraordinary. For example, decoding the human genome took ten years the first time it was done. Now it can be achieved in one week. Companies' investments in Big Data are projected to rise from 19 percent in 2013 to 50 percent by 2016, according to the NewVantage Partners survey.

Big Data has the potential to transform everything.

Don: ***Can you give an example?***

Randy: *One of the impediments to rapid adoption of Big Data solutions in the life science industry is that for each drug, disease and treatment, there are dozens of synonyms to describe each. In the past, there was no easy way to successfully connect these "dots" across databases; and thus, much of the potential discoveries due to cross-pollination never happened. The hurdle the industry must clear in order to connect massive amounts of data is to define a universal way, i.e. a conceptual way, to identify each entity involved, whether it be a chemical, gene, protein or toxic affect. No single software application can address that. A comprehensive approach must be taken.*

Over the past 10 years, semantic languages have been brought to the forefront in an attempt to connect this data, but the fatal flaw is that they can't handle the massive redundancies and ambiguity. The next-level semantic technology is needed, and that's what Code-N is currently launching to the market – a concept-based approach that can connect and interpret this data as well.

Code-N has created the world's largest meta-thesaurus of chemicals, genes, proteins and diseases. Starting with the 2 million chemical-gene-protein terms in the UMLS thesaurus, Code-N added several million more from sources such as DrugBank, KeGG, CAS, HomoloGene, HMDB, ChEBL, MeSH, and dozens of other industry sources to create the most comprehensive

compendium of concepts, synonyms and database identifiers known in the industry. While leveraging this mega meta-thesaurus, Code-N built a series of solutions that can access multiple industry databases simultaneously, whether structured or unstructured, public or private, and connect the-dots between these no matter what the different genes, chemicals or proteins are called. These multiple databases can be queried using simple sentences, with no need for complex syntax or Boolean operators. Now that this concept-based technology is available, exciting advances are possible. Big Data can be leveraged to address current challenges such as competitive surveillance, drug repurposing, advanced safety and toxicology analysis, and resurrecting "shelved molecules."

Don: *Tell me more about how Big Data can help overcome these challenges.*

Randy: *In competitive surveillance, a search for competitive information on compounds, targets and diseases from two companies can simultaneously access 23 million PubMed abstracts, all patent grants and applications, clinical trial data, FDA repositories, and internal databases within 2 or 3 seconds, then send an alert to all interested parties within a company. Even if a competitor is trying to obfuscate its recent research by describing its findings with obscure terms, the concept-based system will be able to identify these and bring them to the light-of-day.*

Big Data can be used to find ways to repurpose drugs coming off patent. This research usually takes weeks or months to find viable opportunities. With new applications that can "connect-the-dots" in Big Data, this process can be made orders-of-magnitude faster. Being able to scour all published literature to find all the targets that "Drug A" affects, these applications can then "follow the trail" to see what disorders can be influenced in a positive way when these targets are affected.

Toxicology analysis benefits from Big Data. When comparing the potential effects of a new compound to known toxic chemicals, scientists are often limited to comparing the up- and down-regulation of just two or three genes and proteins at the same time. The advent of technologies that can more adeptly handle Big Data means there are no longer limits on the amount of comparisons that can be made at the same time. A "digital fingerprint" can be created with the combination of dozens of affects that a new compound is known to cause, and compare that with a library of known toxins within seconds.

And Big Data can also salvage molecules that have been shelved for one reason or another. Every pharmaceutical and biotech company has a catalog or database of molecules or compounds that didn't meet expectations for a specific task. Some were too costly to produce for an intended market or didn't have the expected effect on the particular target. Others were

toxic or simply lost out to a "superior" compound. What if these compounds and their known properties and characteristics could be aggregated in an open-source data store, and shared with the entire industry? This is a perfect challenge for Big Data. Sharing these data pre-competitively would help others avoid making the same mistakes again and again. Imagine being able to search on an idea and quickly see it's a non-starter because the drug class is associated with liver damage. Information like that would make drug discovery smarter and lead the field in more productive directions. Imagine if the billions of dollars "wasted" on failed molecules wasn't wasted, but leveraged to further the science on other discoveries industry-wide.

Don: ***What discoveries did you make that you can attribute to Big Data?***

Randy: *One example of using this breakthrough "concept technology" is a client who was looking to repurpose the cholesterol-lowering drug, atorvastatin (marketed under the name Lipitor®). Since it is now off-patent and has a fairly strong safety profile, they wanted to see how the Code-N solution could speed-up uncovering possibilities they might be able to market. They stated their scientists were able to come up with a new target about once every seven to 10 working days. While using the Code-N repurposing application, they were able to connect-the-dots between 23 million PubMed articles, 1.5 billion pharmacological data statements and 25 years of patents to produce dozens of unclaimed treatment ideas in less than seven seconds. For instance, the data suggested that atorvastatin might be used to treat Huntington's Disease, Multiple Sclerosis (reducing brain plaque), and xanthomas. The data does have to exist to "connect-the-dots" but if it does exist, a conceptual approach can extract these insights whether atorvastatin is called atorlip in one database and totalip, xavator, C33H35FN2O5 or one of its other 96 synonyms in the others.*

Code-N isn't claiming to be an authority on the science of treating diseases – that is left up to the pharma and bio-tech companies it serves. But what Code-N does is provide the informatics solutions that can access massive amounts of Big Data from many disparate sources and bring-to-light ideas within seconds that scientists can then investigate and test.

Another example of a problem that can be solved with Big Data is to be able to quickly identify safety issues with new compounds by creating digital fingerprints of their effects and matching them to a library of known toxins. Identifying those safety issues early can stop development of a compound immediately.

Don: ***What issues or limitations did you have or expect to encounter?***

Marketta: *Disruption in any industry needs to happen when the status quo begins to prevent progress and innovation. It is healthy to challenge inertia and drive change. We've all witnessed frequent reporting on the fact that the life sciences industry is lagging in its ability to innovate. The reasons include slowness and reluctance of converting old infrastructures to new technologies due to risk aversion that drives companies to stick with familiar business models and R&D tools. A scientist at a major conference said: "Unless we learn to value big, potentially disruptive ideas, we won't see transformational breakthroughs." We'll continue to get the status quo: linear, unproductive drug pipelines, siloed data that fails to support open collaboration and partnerships, and outdated regulatory structures and funding models that stifle R&D.*

The good news is that the phenomenon of Big Data is poised to fuel this disruption beneficially. Even though many health care organizations are still wrestling with basic transaction systems, such as electronic medical records, significant initiatives are under way to leverage Big Data to accelerate drug lead discovery, development, repurposing, and safety. There's a strong incentive to leverage all public and precompetitive private data sources to speed up innovation and disrupt the old paradigms. Open collaborative organizations are successfully being introduced to pursue cures for a wide range of challenging diseases. And, on the business side, pharma companies are using Big Data to drive new inventions to make up $35 billion in lost revenue from patent expirations just a year ago.

The life science industry has begun the process of structuring itself more openly. Industry forums and initiatives are valuable contributors. Big Data provides "raw material" for the emerging open environment. New bioinformatics technologies and infrastructures are needed to "cross the chasm."

Of all U.S. industries, the life science industry has one of the biggest gaps in what can be done with Big Data and what is actually being done. Big Data can be leveraged in so many ways to advance science, improve safety and create positive results for patients but we are only in the initial phases of this revolution.

Don: ***Looking ahead to the next few years, what other types of challenges in the life sciences do you expect big data to help address?***

Marketta: *Big Data has become a buzz phrase, which can easily be hyped inappropriately as the long-awaited trove of missing answers and cures. In reality it's the "raw material" for all that and more when enabled through advanced technologies to connect all relevant data*

simultaneously, discard the "noise" and interpret the data in response to "smart" questions. It has the potential of providing deep knowledge for reasoning and predictive capabilities needed for the medicine of the future but not available today.

Today we are dealing with the "first phase" of utilizing Big Data – establishing the baseline for required infrastructures, bridging the silos for instant, simultaneous data access and implementing integrated "intelligent results" via next-generation semantic analytics.

It's quite exciting to think of what lies right ahead of us as a result of the vast Big Data "raw material" and technologies that can extract intelligence from it. Not only will we be able to use real-time data, for example, to repurpose drugs to treat multiple diseases and to discover new warnings of side effects, but also to unlock the power of data to improve clinical care. We'll be able to get to the genome level for real-time detection of diseases and find out in advance whether it's resistant, infectious, etc. We can look forward to relating the entire microbial genome to the human genome and targeting the best treatment, understanding what has to be done to prevent spreading of a disease and how to prevent it in the first place. Future Big Data will assist in "precision medicine" that is expected to use in-depth DNA analysis to personalize drug therapy for patients and search for the genetic causes of diseases.

Don: ***What other Big Data issues are you thinking about?***

Marketta: *The most interesting characteristic of Big Data is that it's heuristic by nature. The information, hypotheses, innovations and feedback that the life sciences industry feeds into Big Data around the clock, increase its "brain power," which in turn enables our concept-based semantic solutions to point out the most up-to-date discoveries to build on, which in turn accelerates next innovations in the industry. We call that the "Heuristic Big Data Cycle!"*