Review Article

IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research

Ying Chen, PhD[1]; Elenee Argentinis JD[2]; and Griff Weber[1]

[1]*IBM Almaden Research Center, San Jose, California; and* [2]*IBM Watson, New York, New York*

ABSTRACT

Life sciences researchers are under pressure to innovate faster than ever. Big data offer the promise of unlocking novel insights and accelerating breakthroughs. Ironically, although more data are available than ever, only a fraction is being integrated, understood, and analyzed. The challenge lies in harnessing volumes of data, integrating the data from hundreds of sources, and understanding their various formats.

New technologies such as cognitive computing offer promise for addressing this challenge because cognitive solutions are specifically designed to integrate and analyze big datasets. Cognitive solutions can understand different types of data such as lab values in a structured database or the text of a scientific publication. Cognitive solutions are trained to understand technical, industry-specific content and use advanced reasoning, predictive modeling, and machine learning techniques to advance research faster.

Watson, a cognitive computing technology, has been configured to support life sciences research. This version of Watson includes medical literature, patents, genomics, and chemical and pharmacological data that researchers would typically use in their work. Watson has also been developed with specific comprehension of scientific terminology so it can make novel connections in millions of pages of text. Watson has been applied to a few pilot studies in the areas of drug target identification and drug repurposing. The pilot results suggest that Watson can accelerate identification of novel drug candidates and novel drug targets by harnessing the potential of big data. (*Clin Ther*. 2016;38:688–701) © 2016 The Authors. Published by Elsevier HS Journals, Inc.

Key words: big data, cognitive computing, data science, drug discovery, genetics, personalized medicine.

Basic science, clinical research, and clinical practice generate big data. From the basic science of genetics, proteomics, and metabolomics to clinical research and real-world studies, these data can be used to support the discovery of novel therapeutics.[1,2] This article reviews a cognitive technology called IBM Watson and describes early pilot projects. The project outcomes suggest that Watson can leverage big data in a manner that speeds insight and accelerates life sciences discoveries. This commentary is a 5-part discussion of the following: (1) the need for accelerated discovery, (2) the data hurdles that impede discovery, (3) the 4 core features of a cognitive computing system and how they differ from those of previous systems, (4) pilot projects applying IBM Watson to life sciences research, and (5) potential applications of cognitive technologies to other life sciences activities.

PART I: SOLUTIONS THAT CAN ANALYZE BIG DATA ARE NEEDED IN LIFE SCIENCES RESEARCH

Although debated, recent estimates suggest that the costs of bringing a new drug to market has reached $2.5 billion and >12 years of investment.[3] Of drug candidates, 80% to 90% fail to gain U.S. Food and Drug Administration approval.[4] The most common reasons for failure include lack of efficacy, lack of safety, poor dosage selection, and poor endpoint selection.[4,5] Looking across disease states, in some therapeutic areas approval rates are as low as 6.7%.[6]

Accepted for publication March 1, 2016.
http://dx.doi.org/10.1016/j.clinthera.2015.12.001
0149-2918/$ - see front matter

Changing market dynamics have increased the hurdles to developing a successful drug candidate. Greater availability of generic drugs is one of these hurdles. Generic prescriptions made up 82% of all prescriptions dispensed in 2014.[7] In established therapeutic areas such as cardiovascular disease, studies have compared generic drugs to brand-name medications. The study results indicate that generic drugs are associated with an 8% reduction in cardiovascular outcomes including hospitalizations for acute coronary syndrome, stroke, and all-cause mortality versus their brand-name counterparts. These outcome improvements were attributed to better adherence to generic drugs.[8] In type 2 diabetes, hypertension, and hyperlipidemia there are 4 or more generic agents available.[9] Brand-name medications in these therapeutic areas come with costs of 8 to 10 times more than costs of available generics. New agents must therefore be more effective or safer than existing low-cost generic options to justify the price differential. Biopharmaceutical companies have shifted to orphan (rare) diseases and cancer in which there is a dearth of medications and regulatory incentives.[10] Orphan diseases also present challenges to the discovery of new therapeutics including the heterogeneity of diseases, inadequate understanding of the natural history of these diseases, and a lack of biomarkers to aid outcomes studies.[11] In both cases, companies need to speed advances in research. They need to accelerate breakthroughs in disease pathophysiology, drug target identification, and early candidate evaluation so that more viable drug candidates can be approved and provided to patients.

Today's life sciences researchers have a spectrum of data available to them. The data come in many varieties including high-throughput screening, genomic sequencing, mass spectrometry, metabolomics and transcriptomic data, phenotyping, and more.[12,13] Big data are important to scientific discovery because they hold the potential to unlock the origins of disease and open up new avenues to prevention and treatment. For example, gene sequencing, which helps to identify gene mutations that cause diseases, generates terabytes of data. This dataset alone is a challenge to maintain in a sustainable environment while also allowing rapid analysis of the information. The data quickly become unmanageable when you add other types of data such as proteomics and metabolomics. If these datasets could be combined and connected to other data types, insights from each dataset could be pieced together to unlock understanding about the origins and processes of many diseases.[12] The challenge lies in combining, interpreting. and analyzing vast and disparate data types from different sources.[12] As a result, the potential of big data has yet to be fully realized because current solutions cannot fully contend with its scale and variety. [12,13] There is a need for technology solutions that address these issues to enable more productive and efficient research. Ultimately, the benefit that these technologies should confer is the accelerated discovery of viable drug targets, drug candidates, and other novel treatment modalities.

PART II: THE TYPES OF CHALLENGES POSED BY BIG DATA

Big data come with inherent challenges that include volume, velocity, variety, and veracity. Each of these facets should be addressed in the design of a technology solution. First, a solution must be able to manage the sheer volume of data available and "keep up" with integrating new data that are constantly being produced. There are nearly 200,000 active clinical trials, 21,000 drug components, 1357 unique drugs, 22,000 genes, and hundreds of thousands of proteins.[14,15] Each of these areas of study includes testing and experiments that yield vast quantities of data, making it difficult for any 1 researcher or even teams of scientists to absorb.[15–17] There are >24 million published medical and scientific articles in the 5600 journals in MEDLINE alone, with 1.8 million new articles published annually.[18,19] Meanwhile, the average researcher typically reads on average 250 to 300 articles in a given year.[20] This suggests that scientists may not be keeping up with the basic science published in their area of specialty, let alone making novel connections that could come from harnessing many data sources.[20,21] The volume of published science grows at a rate of $\sim$9% annually, doubling the volume of science output nearly every 9 years.[22] The ability to absorb only a fraction of available information results in many lost opportunities to further research. Drug discovery depends on identifying novel and effective targeting strategies that produce better clinical outcomes for patients. Harnessing volumes of information about how disease processes originate and progress and how drugs affect animals and humans could yield novel treatment strategies.

In order to unlock the potential in data, data must be understood in all its varieties. Structured data

include data contained in tables or data cells such as names, addresses, and isolated lab values. Today, a high percentage of data is unstructured. Unstructured data are information such as text where meaning is often derived from context. Other unstructured data types include images, X-rays, sonograms, electrocardiograms, magnetic resonance images, and mass spectrometry results.[13]

Data variety is often accompanied by the issue of data silos. For example, with respect to any biopharmaceutical company's drug, data from chemical reaction experiments, high-throughput screening, animal studies, human trials, and postmarketing drug safety surveillance are often kept in different repositories. Data silos exist in most organizations, and existing approaches to integration and analysis have not been completely successful in addressing data scale or diversity. Additionally, there are hundreds of external data sources covering patents, genes, drug labels, chemical compounds, proteins, and published studies. For a solution to successfully address these challenges, it must be able to support the aggregation of big data from multiple sources and retain these data in a stable, secure environment with efficient processes for integrating new data so that the insights generated are accurate, current, and relevant.

Another challenge with large datasets is the presence of data "noise." The term noisy data refers to information that is dense, complex, or characterized by conflicting indicators that can make drawing confident conclusions from it difficult. Conflicting and "noisy" data are a common issue in most fields including life sciences and medicine. This issue is particularly important in medicine in which evidence-driven decisions are the foundation for caring for patients. Study veracity, quality, and replicability are often under discussion.[20] Resolving evidentiary conflicts or at least surfacing them while pointing directly to the publication passages would offer researchers the opportunity to read the source text and evaluate the information further. Today's systems tend to rely on humans to curate (collect and organize) and evaluate evidence, which presents 2 problems. First, a human may not encounter the evidence in favor of or against a particular hypothesis if it is buried in millions of pages of data. Second, humans tend to approach problems with some bias. A cognitive system could access more data and surface any evidence in favor of or against a hypothesis.

PART III: COGNITIVE TECHNOLOGIES: A NEW WAY TO AGGREGATE AND UNDERSTAND BIG DATA

Cognitive technologies are an evolution in computing that mimics some aspects of human thought processes on a larger scale. In this case, scale refers to the ability to process the volumes of data and information available in the scientific domain. Technology developers have realized that human reasoning, learning, and inference comprise one of the most sophisticated thinking systems in existence.[23–25] Still, human cognition has limitations, 2 of which include scalability and bias. Cognitive systems attempt to mimic aspects of human thinking while adding the ability to handle large amounts of information and evaluate it without bias.

In the computing community, the definition of cognitive computing is a topic of debate. It is often associated with artificial intelligence (AI), a field of technology that covers broad aspects of human intelligence.[26] AI includes the skills related to reasoning and problem solving but also perception (face recognition and vision) and the ability to manipulate objects (robotics).[26] In this paper, cognitive computing refers to a combined subset of these technologies that read, reason, learn, and make inferences from vast sets of unstructured content.[27]

Even in the area of cognition, AI tends to focus on individual algorithms and models that mimic specific human cognitive functions (ie, reading), whereas the cognitive computing solution described in this paper is a holistic system in which the competencies of reading, reasoning, and learning are grouped together to answer questions or explore novel connections.[27] Some aspects of cognitive computing, such as the ability to address data volume, velocity, variety, and veracity, are not areas of focus in the AI development community. Cognitive technologies are needed because they address data challenges by applying multiple technologies to enable comprehension of vast, disparate data sources in a single solution. Through a comprehensive approach to data aggregation, comprehension, and analysis, along with technologies that read, reason and learn, more novel avenues in research could be discovered.

To understand how cognitive computing works, it is helpful to compare and contrast how human beings and cognitive technologies engage in discovery and various forms of decision-making processes. One way to describe these processes is observation, interpretation, evaluation, and decision.

Observation of data is the first step in creating a cognitive system. It refers to the aggregation, integration, and examination of data as a foundation for evaluation and discovery. Humans observe through different sensory channels, such as reading relevant publications or listening to others. Humans also often have a pre-existing foundation of information gained through their own observation, education, and life experiences. These observations are retained in memory as part of a broader knowledge base.

In order to make observations, a cognitive solution requires access to volumes of data. The identification, purchase, licensing, and normalization of data must all be coordinated. With a cognitive computing system, hundreds of external, public, licensed, and private sources of content that may contain relevant data are aggregated. In the case of Watson, IBM aggregates these data into a single repository called the Watson corpus. A unique Watson corpus is established for each domain to which Watson is applied. Therefore, in law, medicine, engineering, and finance, a tailored Watson corpus could be created with datasets and content relevant to that domain. The content is normalized and cleansed into a formatted dataset that can be used for analysis.

Interpretation entails the ability to understand data, in this case, language beyond the definitions of individual terms, to deduce the meaning of sentences and paragraphs. As humans, we take in content and information. We read and recognize words, translating them into abstract meaning. For example, a chemist will recognize compounds from published articles or patents and create a mental representation of related compounds and the features that define them.

Similarly, a key component of a cognitive system entails learning the language of a specific industry or domain. To enable language comprehension, a system must be supplied with relevant dictionaries and thesauri. These might include a known list of human gene names or chemical names, but they also include the verbs that create relationships between them such as "express or inhibit." Understanding the verbs, nouns, and prepositions in each sentence makes cognitive systems different from key word search and text analytics that may identify only the nouns of interest or rely on matching individual words to find relevant information. The ability to understand verbs, adjectives, and prepositions enables comprehension of what language means versus just what it says.[26]

Figure 1 shows how a system like Watson is taught how to recognize and reconcile multiple names or synonyms for an entity into a single concept. A cognitive system that learned about chemistry would recognize Valium as a chemical structure. It will not only recognize Valium, but also resolve >100 different synonyms for Valium into a unique chemical structure. An investigation into any chemical will find relevant documents that contain different forms of that chemical's name, not just its brand name, for example (Figure 1). This capability is an inherent part of a cognitive system.[28,29] The interpretation of other data formats like magnetic resonance images, echocardiograms, or any other visual data should be contemplated in future solution iterations.

Like humans, a cognitive system can leverage known vocabulary to deduce the meaning of new terms based on contextual clues. A chemist can recognize a newly discovered compound because it shares attributes with other compounds that he or she has seen before. Similarly, a cognitive system can identify a newly approved drug by recognizing contextual clues like a discussion of its indication or side effects. This learning ability is one of the greatest differentiators between cognitive and noncognitive technologies. In domains such as life sciences, in which new diseases, drugs, and other biological entities are continuously being discovered, solutions that rely on humans to manually update their knowledge base could miss important insights.

Once relevant datasets are collected and Watson has been provided with dictionaries that enable it to recognize terms, a set of *annotators* is applied to the data. These annotators extract specific nouns and verbs that convey relationships between proteins (Figure 2). A chemical structure annotator will be able to extract International Union of Pure and Applied Chemistry or chemical names and convert them into unique chemical structures out of the text of scientific journal articles.[29] Similarly, a gene or protein

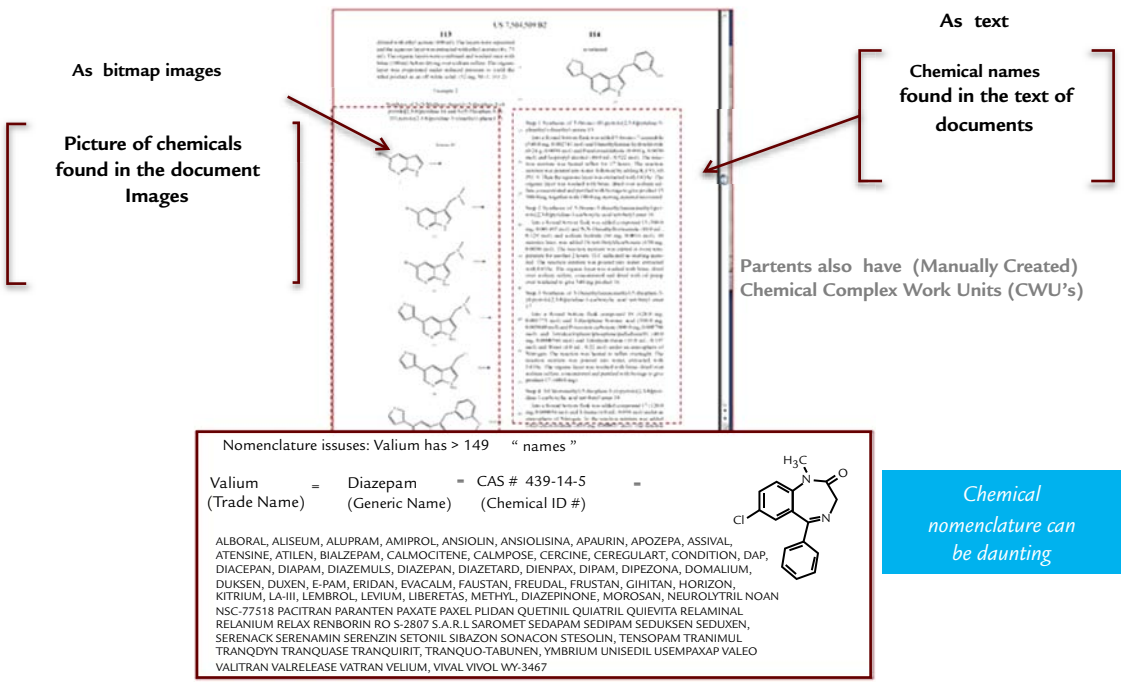

Figure 1. Example of Watson concept recognition. Watson can recognize terms and many of their synonyms. Watson can recognize a chemical such as Valium whether expressed as a chemical diagram, smile string, chemical identification number, its generic name, and other synonyms.

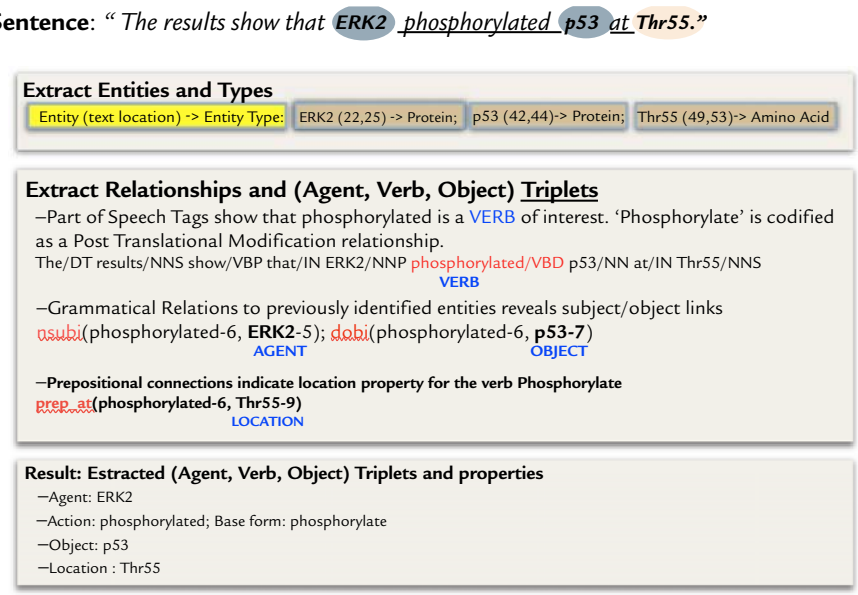

Figure 2. Sample of Watson extracting entities. A cognitive system like Watson uses annotators which are specialized types of computer code to read and extract terms from scientific literature. Watson reads the sentence and recognizes nouns as proteins, verbs like phosphorylation, the object of those verbs and the location of an event.

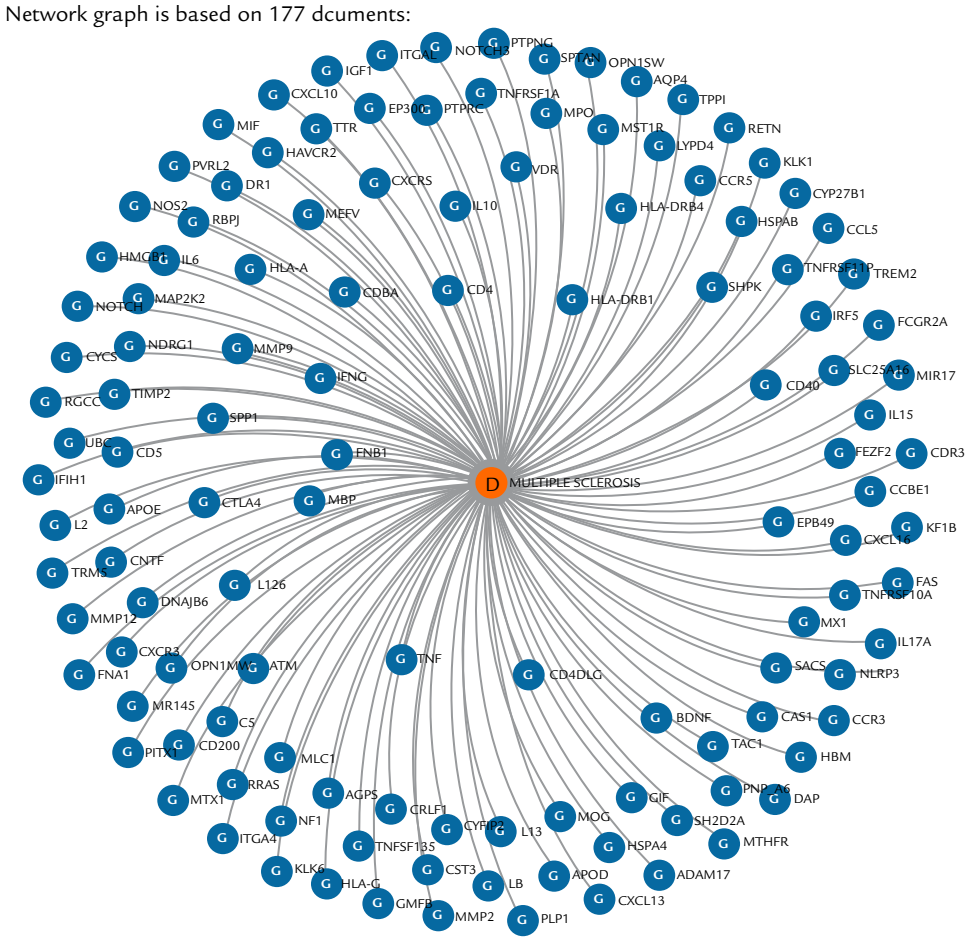

Figure 3. Watson depicting gene relationships to multiple sclerosis. This is an example of a network of genes that Watson has produced in real time when a user types in a disease name. Watson generates these network maps in real time, using its annotators to extract the relationships between any gene and multiple sclerosis out of > 26 million MEDLINE abstracts.

annotator can extract gene and protein names and resolve gene synonyms to a single unique gene entity. In addition to extracting individual entities such as genes, Watson's annotators identify the relationships among genes, drugs, and diseases.

These annotators typically learn from patterns in the text where they occur and then extrapolate more generally for a given type of entity. For example, if a gene annotator sees examples of "P53, WEE1, and ATM" in the context of MEDLINE journal publications, it will apply machine learning and domain rules to "figure out" other words and phrases in the text that look like the genes *IL17*, *IL1*, and so on.[30,31]

Deep natural language-processing and machine-learning techniques were developed so that Watson could teach itself about these concepts and comprehend the subject at a more meaningful level. Figure 3 is an example of a life sciences annotator that extracts protein relationships from unstructured text. Figure 3 illustrates how major components of the sentence are processed so that the protein ERK2 is recognized as the acting agent because the next word "phosphorylates" is recognized as a verb along with the object of that verb, P53. Deep natural language comprehension will also understand a preposition such as "on" as a location with a trigger in the

computing code to extract the site of that location (in this case, therine 55). In this specific example, the recognition and extraction of these parts of language enables a technology to recognize a relationship between 2 kinases, the type of relationship, and the location of their interaction.

When we apply domain annotators on large volumes of unstructured content via the IBM Watson technology, its processing speed extracts chemicals, genes, and drugs from thousands of scientific publications and patents within hours. Extraction of this information by humans would likely take significantly longer. The information extracted by annotators from millions of pages of text can then be composed into a wide variety of runtime analytics and visualizations. Figure 2 shows 1 example of the visualizations that one can create on top of annotated datasets.

In addition to visualizing expressed patterns, machine learning and predictive analytics generate hypotheses about relationships; in effect "inferring" novel connections not yet supported by overt statements in the literature as was done in a project with Baylor College of Medicine where hypotheses of new kinases that could phosphorylate TP53 were generated out of existing medical literature.[31,32]

If the observation and interpretation of concepts are the foundation for discovery in a human's cognitive process, the next step is evaluation.[30] Humans have the ability to evaluate evidence and apply it to solve different types of problems. Evidence can be evaluated to provide a single best evidence-based answer to a query or offer several answer candidates. In the case of research, Watson evaluates evidence for the purpose of exploration and discovery. Watson uses evidence from text to compose visual networks of data such as all the genes with a relationship to Alzheimer's disease. In this case, holistic view refers to composing a visual depiction of all the items in a specific group and their relationships to each other. In the case of Watson Discovery Advisor, evidence is not evaluated to come to a "best" answer. It is used to discover novel relationships and new hypotheses for further evaluation.

Once a cognitive system gains basic understanding of domain terminology, it translates fragmented information from content into holistic pictures using various visualizations and summarization techniques. For example, a researcher looking for new potential drug targets in multiple sclerosis (MS) may want to see evidence of any genes with a relationship to that disease. Figure 2 illustrates a network map composed by Watson in an attempt to compose a holistic view of the genes associated with MS. In less than a minute, Watson processed 24 million MEDLINE abstracts and hundreds of other pages of content and found 177 documents mentioning genes with a connection to MS. More importantly, gene relationships are depicted so that a user can see their relationship to MS and to each other without having to read all 177 articles. If a user right clicks on any connecting chord between 2 genes, the relationship between them is summarized (ie, positive regulation) and the researcher can access the specific passage in the article describing the relationship. Cognitive technologies analyze and create holistic network maps at run time, meaning that the visualization is created at the time that the user makes the request. Humans would need to read each one of the 177 documents and manually draw these maps and then update them each time new information is published. This manual processing usually means that there is a delay between the time that the data become available and the time that the data are incorporated into any solution relying on them. Cognitive systems automatically update their visualizations when provided with new content, which enables researchers to leverage the most current data from which to make new discoveries.

In the context of discovery, the term *decision* refers to the ability to make a determination or take action based on data. Such determinations may be to conclude that a specific protein may be a worthy new drug target to validate in experiments. Decisions in life sciences and medicine rely heavily on evidence. Watson helps support confident decisions about where to focus research by making evidence readily accessible to the researcher. Further, Watson leverages quantitative predictive analytics to infer relationships for which there may not yet be explicit evidence. In this case, Watson relies on a researcher to provide a known set of items like genes with a relationship to a disease. Watson uses those known genes to train itself to identify other genes with like text traits. Researchers then provide Watson with a candidate list. The candidate list is a group of genes, diseases, or drugs that a researcher would like to narrow down to a list of high-potential options for further testing. Watson's ability to score a list of candidates by

their text features may help researchers accelerate their research by focusing on those Watson ranks highest.

Presenting a Holistic View

In order to create a list of potential gene, drug, or disease candidates that a researcher decides is worthy of taking to experiments, he or she must first explore and identify novel relationships. Watson combines its ability to observe, interpret, and evaluate with novel data representation approaches. Two of the more novel approaches include presenting information in holistic relationship maps and promoting cross-domain linking. Holistic relationship maps are visual networks of relationships that help researchers see a full depiction of a group of drugs, diseases, and genes and their connections. Cross-domain linking refers to leveraging data across all therapeutic areas, study types, and disciplines to inform those novel connections or hypotheses.

As discussed previously, much of the data generated and collected both publicly and privately about a given drug are kept in silos. As a result, what might have been learned during different phases of development is often isolated from other insights. Additionally, the tools used to analyze these data are often specifically designed for use at particular phases of drug development such as mass spectrometry for protein identification. During drug design, development, and clinical research, different groups of researchers each use unique data types and analytical tools to understand data that are maintained specifically for their type of research. Cognitive solutions enable the integration of these data to combine discoveries made across the drug's development life cycle. Watson then creates a holistic visualization out of all the data in simple formats such as network maps, depicting more relationships and a fuller view of the information available.

From Serendipity to Intention: Making Cross-Domain Linkages a Core Feature Discovery

Harnessing big data and presenting them in holistic visualizations can encourage the identification of novel connections that would otherwise have only been made by chance. Science and medicine have experienced leaps forward through fortuitous discoveries. Well-known examples include penicillin, cisplatin, warfarin, and the smallpox vaccine.[33–35] Serendipity has been attributed to up to 24% of all discovered drugs.[34,35] Often the discovery emerged from connections made when seemingly dissimilar domains were brought together by chance. For example, an observation that people resistant to the smallpox disease were often dairy farmers in close contact with infected cattle eventually led to the creation of an effective vaccine for humans.[33]

Cognitive computing technologies can be configured to make cross-domain linkages versus rely on serendipity. For example, insights about a gene entity such as its role or function can be derived from many sources. A cancer researcher seeking to discover the role of a given gene in cancer may miss insights if the search is limited to the literature related to that disease. A cognitive discovery platform will surface all information about a given entity regardless of the disease, journal, or even species of study.

Another way in which cognitive discovery uses cross-domain linkages is demonstrated in drug repurposing. Big data could be useful in drug repurposing because information about drugs, their mechanisms of action, targets, effects, and outcomes could be used to inform development of new therapies. The challenge is that data about drugs are kept in a variety of repositories such as the animal study results from preclinical studies, clinical trial data generated from Phase I through III clinical trials, the labels of all approved therapies, and the adverse event reports kept in drug safety databases. In this case, Watson can be used to look across all of this information, exploring all drugs for mechanism-of-action similarity or across all diseases for shared pathways such as an inflammatory pathway or an immunological pathway. A drug label, animal study, *in vitro* cell experiment results, and human trials combined may reveal a novel relationship that could help unlock a new indication. One of the "test" cases for Watson Discovery Advisor discussed later in this article further discusses the use of Watson for drug repurposing in the treatment of malaria.

PART IV: HOW WATSON DISCOVERY ADVISOR COULD AID LIFE SCIENCES RESEARCH

Figure 4 illustrates the basic architecture for Watson as it has been applied to the life sciences domain. The

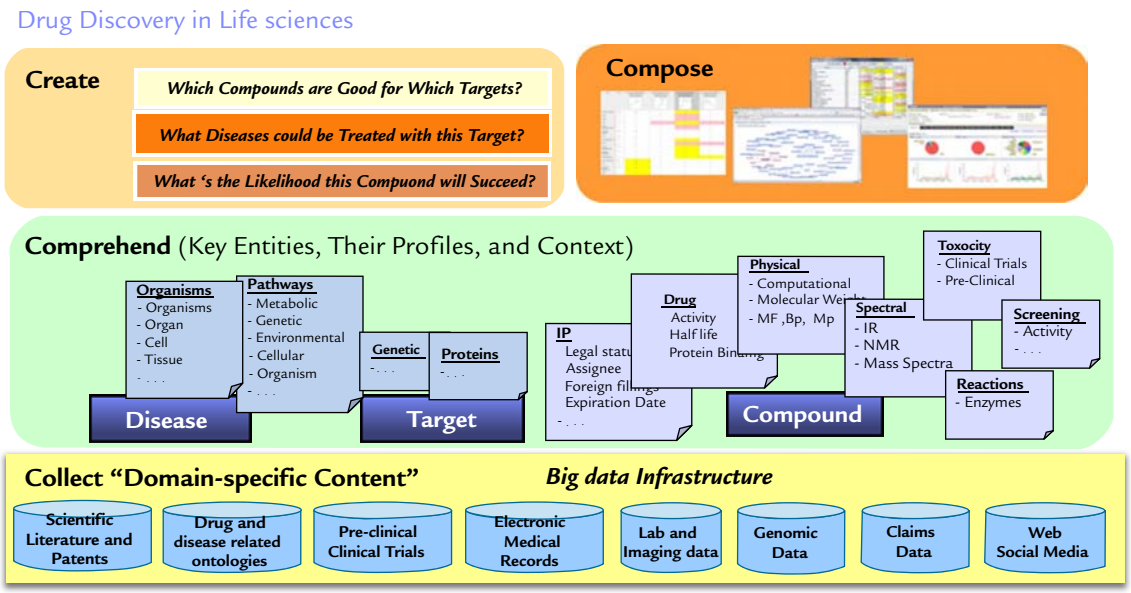

Figure 4. Watson Discovery Advisor applied to Life Sciences. This is a depiction of the architecture Watson leverages when being applied to the life sciences domain including the aggregation of data, the recognition of major scientific concepts and then an interface that enables a researcher to explore.

figure illustrates the layers that comprise a cognitive solution starting with the content and data sources that are relied on by researchers. In addition to these data, Watson is trained to recognize and understand the key concepts such as recognizing genes and the role they play or drugs and their relationships to indications and side effects. The top layer is the dynamic user interface that surfaces graphics like bars showing posttranslational modifications on the sites of a protein, for example, or a group of articles retrieved in response to a query. As shown in Figure 4, the foundation of a cognitive system involves aggregating different data types. By aggregating big datasets, Watson is then in a position to find connections between them. Researchers are challenged to replicate this output because they often cannot gain access to such volumes of data and lack technologies to pull the data together and find the meaningful connections. Today, there are some tools that attempt to do this, but the data are often manually searched, reviewed, and mapped by humans, which limits the amount of data and number of sources that can be leveraged. There are also limits to the speed at which the data can be evaluated and that introduce human bias into the discovery process. In life sciences, examples of the relevant data types include published literature, patents, preclinical animal study reports, clinical trial data, genomic information, drug labels, and pharmacology data. Some data can be accessed from public domain or third-party content providers. Other content might be owned by private enterprises. Some data will come in structured data formats such as chemical structures, whereas others might be unstructured, such as published journal articles. A solution should be built to ingest and link such datasets in a secure, coherent, and scalable fashion and anticipate continuous creation of new data, new data formats, and emergence of new data types.

Today, Watson Discovery Advisor for Life Sciences is supplied with dictionaries, thesauri, and ontologies on genes, proteins, drugs, and diseases. It includes annotators that are tested for accuracy in recognizing, extracting, and categorizing these entities. Last, the data are surfaced for evaluation via network maps, co-occurrence tables, and other visualizations that promote insight. These visualizations along with a question-and-answer function allow researchers to seek evidence-supported answers to a question or explore any relationship for which any level of evidence exists.

Watson's Life Sciences Use Cases

IBM Watson was originally a research experiment to determine whether a computer could be taught to read volumes of text such as Wikipedia, newspapers, and other text-based sources of information and produce reliable evidence-driven answers in response to natural language questions. The project culminated in a public demonstration of the technology on the game show Jeopardy in 2011 during which Watson defeated 2 human Jeopardy champions. Shortly thereafter, several organizations from different industries approached IBM to understand where Watson could be adapted to specific business challenges. Baylor College of Medicine was one such group. Watson Discovery Advisor has been applied in several pilot projects. Two of them described here include a study of kinases in cancer and repurposing of drug compounds for potential treatment of malaria.

Test Case 1: Baylor College of Medicine: A Retrospective and Prospective Exploration of Kinases

In 2013, Baylor College of Medicine approached IBM to understand whether Watson Discovery Advisor could enhance insight into cancer kinases. The research project included a retrospective and prospective exercise exploring kinase relationships.

Research Question

The research challenge was whether Watson could predict kinases that might phosphorylate the P53 protein. In order to evaluate Watson's predictive abilities, a study had to be designed in which Watson used information from the past to identify kinase relationships that were later validated and published. For the pilot, Watson was trained on kinases that had been observed to phosphorylate P53 using evidence in published literature through the year 2002. Watson then used this training and the MEDLINE abstracts through 2002 to "guess at" the kinases discovered to phosphorylate P53 in the following decade spanning 2003 to 2013.[31]

Study Design

In designing the retrospective experiment, Watson was provided a set of human kinases known to phosphorylate P53. Using text mining, Watson read all the articles discussing the known kinases provided to IBM. With text feature analysis and graph-based diffusion, Watson found and visualized text similarity patterns between these kinases. Once these models were refined, they were applied to the MEDLINE abstracts up through 2002 to determine whether Watson could identify the kinases discovered in the period 2003 to 2013. Figure 5 illustrates how the kinase relationships were mapped based on their literature distance. The kinases whose text patterns were most similar to a set of kinases already known to phosphorylate P53 suggested that they had highest likelihood of also phosphorylating P53.[31]

RESULTS

The results of the study were that over the course of several weeks, IBM researchers using Watson technology identified 9 potential kinases that would phosphorylate P53. Of these, Baylor validated that 7 had in fact been discovered and validated through published experiments during the following decade from 2003 to 2013.[31] These results suggested that cognitive computing on large unstructured datasets could accelerate discovery of relationships between biological entities for which there was yet no explicit evidence of their existence.

Watson was then applied in a prospective exploration of the full set of MEDLINE abstracts through 2013. Watson surfaced a ranked set of additional kinases with various levels of probability of phosphorylating P53. PKN1 and NEK1 were 2 kinases highly ranked by Watson as having the potential to phosphorylate P53.[32] Lab experiments at Baylor College of Medicine suggested that these kinases could phosphorylate P53 in both in vitro experiments and in tests on human cells. Further experiments are being conducted to test the activity of these kinases in organisms. The results of the retrospective study were published in the 20th Association for Computing Machinery (ACM) Society for Knowledge Discovery in Data SKDD international conference proceedings held in August 2014.[31] The prospective Watson study identifying PNK1 and NEK1 was published in August 2015.[32]

Test Case 2: Applying Watson to Drug Repurposing

Cross-domain discovery, the detection of a new insight or relationship based on information from 2 or more domains, was demonstrated in a pilot project with

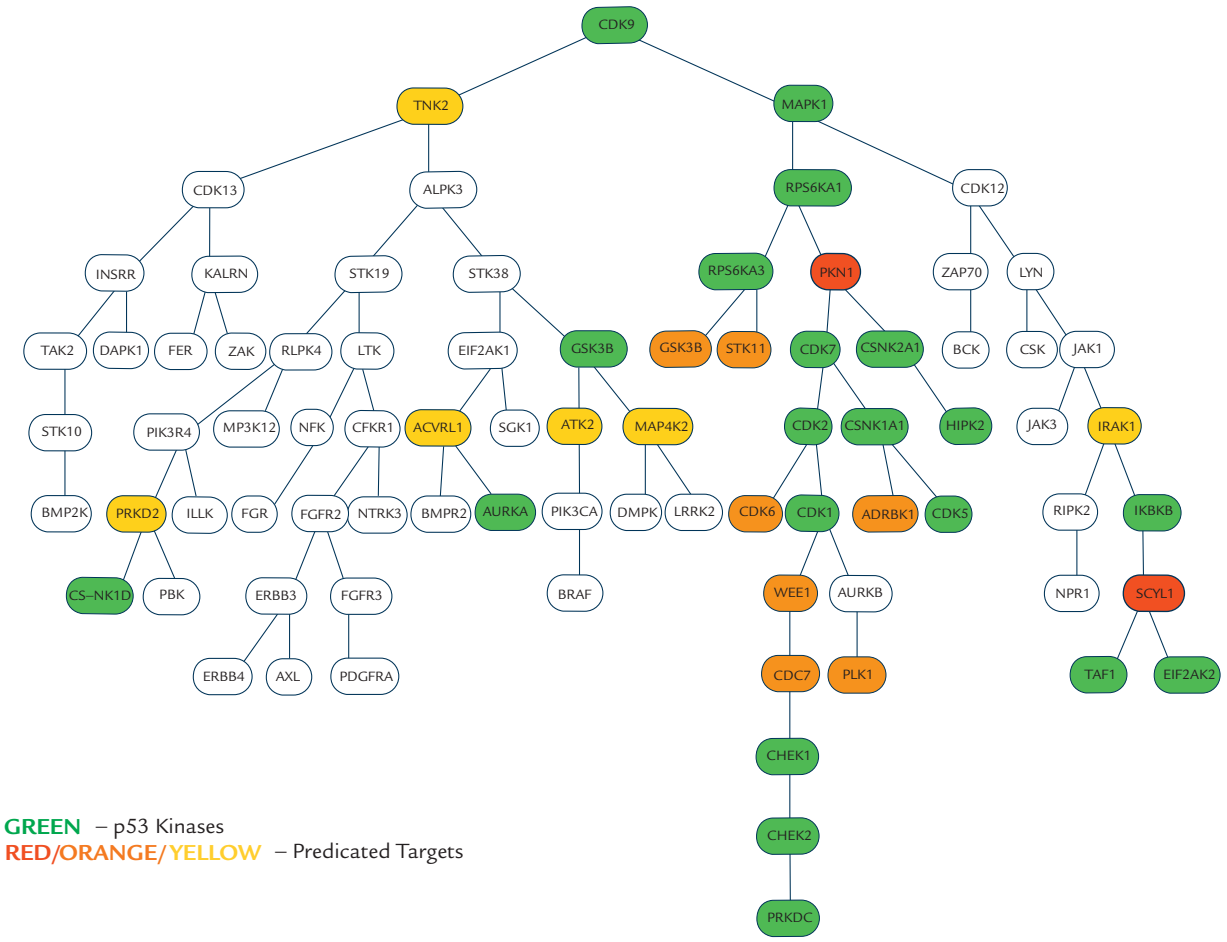

Figure 5. High-potential cancer kinase network. This is the result of Watson performing text analytics on kinases known to phosphorylate PT53 and a group of proteins that were tested to evaluate their potential to phosphorylate TP53.

a large biopharmaceutical company. In this case, the project objective was to identify compounds in the company's existing therapeutic portfolio with the potential to treat of malaria. The study method included exploration of MEDLINE literature looking across all drugs approved for use in humans. Watson then searched for statements suggesting efficacy against the malaria parasite. The second part of the study method looked at all of the company's existing compounds and identified any that had a structural similarity to known malaria treatments by looking for similarity in chemical structure and mechanism of action.

The result of this proof-of-concept project was that 15 drug candidates from the company's existing portfolio were identified for further study.[36] The actual duration of the exploration using Watson technology took less than 1 month. The company had been working on this endeavor with at least 10 research scientists over 14 months and had found a similar number of candidates. Between the 2 lists of candidates generated by the pharmaceutical company and Watson, about half of them were the same. The other half of candidates on the list produced by Watson were candidates that the company researchers had not identified during the course of their research. The company chose to take the results internally, and further study of the candidates was never disclosed.

PART V: THE FUTURE OF COGNITIVE DISCOVERY

Early pilot research projects with Watson in cancer kinase research and drug repurposing suggest that the

attributes of a cognitive system could potentially aid researchers in making connections out of large datasets faster than and potentially aid them in making connections that they may not have otherwise considered.[31,32,36] To determine where cognitive systems might add the most value, Watson should be applied to a breadth of research questions. Although Watson has been applied to research on cancer kinases and drug repurposing, other projects such as predicting combinations of genes or proteins that as a group may play a role in disease onset or progression should also be attempted. The projects should cover a breadth of entities from biomarkers to biological processes to biologics and should cover several therapeutic areas to determine whether the predictive models can be used across disease states. Exercises using various data types will also yield important information about whether predictive models can be further enhanced by combining structured and unstructured data to unlock novel insights. If Watson can be successfully trained on a breadth of entity types across disease states, it could help accelerate discoveries about disease origins, contributing pathways and novel drug targets.

Additionally, the current capability of Watson to read and extract relationships from text is being applied to pilot research projects in pharmacovigilance. A few research projects with large pharmaceutical companies have involved the application of Watson to reading both published journal articles and adverse event case reports to evaluate whether Watson can assist the drug safety process through faster recognition and coding of adverse events out of text. In this case, Watson may be used to augment existing drug safety personnel to speed their work and support timely reporting of adverse events to US and European regulatory agencies.

CONCLUSIONS

Cognitive computing solutions patterned after several key aspects of human thought are emerging in many industries. Their ability to ingest varieties of data and to understand, evaluate, and learn from the data has the potential to unlock novel insights. These solutions may enhance areas such as Life Sciences, which are in dire need of accelerated innovation. Early pilot projects discussed here suggest that cognitive computing infuses novelty and adds speed to the research process. Further study is needed to validate its utility in different therapeutic areas and research domains. Cognitive computing may also add value in the identification and coding of adverse event reports from the text of case reports and published articles. Current pilot projects are beginning to yield insight into whether Watson has the potential to improve both the accuracy and speed of adverse-event detection and coding. As with discovery, multiple test cases across event types, drug types, and diseases will be needed to evaluate and improve Watson's abilities in drug safety. In both cases, IBM will learn from each engagement and improve Watson's ability to both extract known relationships and hypothesize novel relationships through predictive text analytics.

ACKNOWLEDGMENTS

No other authors or writers participated in the writing of this article. Elenee Argentinis, corresponding author made all revisions, referencing and and submitted all versions as well as coordinated the creation of all graphics. Ying Chen, lead author and technical lead was responsible for the description of Watson technology and cognitivie computing. GRiff weber obtained client permissions to discuss projects, ensured factual accuracy of client project descriptions.

CONFLICTS OF INTEREST

The authors have indicated that they have no other conflicts of interest regarding the content of this article.

REFERENCES

1. Raghupathi Wullianallur, VIju Raghupathi. Big Data Analytics in Healthcare: Promise and Potential. *Health Inf Sci Syst*. 2014;7:3.
2. Holzinger Andreas, Dehmer Matthias, Jurisica Igor. Knowledge discovery and interactive data mining in bioinformatics—state-of-the-art, future challenges and research directions. *BMC Bioinformatics*. 2014; 15(Suppl 6):I1.
3. Avron Jerry. The 2.6 Billion Dollar Pill: Methodologic and Policy Considerations. *N Engl J Med*. 2015;372:1877–1879.
4. Kola Ismail, Landis John. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat Rev Drug Discov*. 2004;3:711–715.
5. Sacks LV, et al. Scientific and Regulatory Reasons for Delay and Denial of FDA Approval of Initial Applications for new Drugs 2000-2012. *JAMA*. 2014;311:378–384.

6. Hay Michael, Hala H Shamsuddin, Yasinskaya Yuliya I, et al. Clinical Development Success Rates for Investigational Drugs. *Nat Biotechnol*. 2014;32:40–51.
7. Keehan Sean P, Cuckler, Gigi A, et al. National Health Expenditure Projections, 2014-24: Spending Growth Faster than Recent Trends. *Health Aff (Millwood)*. 2015;34: 1407–1417.
8. Gagne Joshua J, Choudhary, Niteesh K, et al. David. Comparative Effectiveness of Generic and Brand-Name Statins on Patient Outcomes. *Ann Intern Med*. 2014;161: 400–408.
9. Shrank William H, Choudhry Niteesh K, Liberman Joshua N, Brennan Troyen A. The Use of Generic Drugs in Prevention of Chronic Disease is Far more Cost-Effective than Thought, And May Save Money. *Health Aff (Millwood)*. 2011; 30:1351–1357.
10. Meekings Kiran N, Williams Cory SM, Arrowsmith John E. Orphan Drug Development: an Economically Viable Strategy for Biopharma R&D. *Drug Discov Today*. 2012;17: 660–664.
11. Melnikova Irena. Rare Disease and Orphan Drugs. *Nat Rev Drug Discov*. 2012;11:267–268.
12. Bellazzi Riccardo. Big Data and Biomedical Informatics: A Challenging Opportunity. *IMIA Yearbook of Med Inf*. 2014:8–13.
13. Higdon Roger, Winston Haynes, Stanberry Larissa, et al. Unraveling the Complexities of Life Sciences Data. *Big Data*. 2013;1:42–50.
14. Overington John P, Al-Lazikani B, Hopkins A. How Many Drug Targets Are There? *Nat Rev Drug Discov*. 2006;5:993–996.
15. Wu, Po-Yen, Venugopalan, Janani, Kothari, Sonal, et al. April, 2014. Big data analytics – biomedical and health informatics for personalized cardiovascular disease care. Lifesciences.ieee.org. http://lifesciences.ieee.org/publications/newsletter/april-2014/539-big-data-analytics-biomedical-and-health-informatics-for-personalized-cardiovascular-disease-care. Accessed May 26, 2015.
16. Chen Philip, Zhang CL, Chun-Yang. Data-intensive applications, challenges, techniques and technologies; a survey on big data. *Inf Sci*. 2014; 275:314–347.
17. National Institute of Health. Trends, charts and maps. https://clinicaltrials.gov/ct2/resources/trends. Accessed August 9, 2015.
18. NIH U.S. National Library of Medicine MEDLINE® Fact Sheet. https://www.nlm.nih.gov/pubs/factsheets/medline.html. Accessed August 9, 2015.
19. Eveleth R. Academics write papers arguing over how many people read (and cite) their papers. Smithsonian.com. 25, 2014. http://www.smithsonianmag.com/smart-news/half-academic-studies-are-never-read-more-three-people-180950222/?no-ist. Accessed May 26, 2015.
20. Van Noorden R. Scientists may be reaching a peak in reading habits. February 3, 2014. http://www.nature.com/news/scientists-may-be-reaching-a-peak-in-reading-habits-1.14658. Accessed May 26, 2015.
21. Eveylth, Rose. Smithsonianmag.com. Smithsonian.com, March 25, 2014. Accessed February 15, 2016.
22. Van Noorden, Richard. May 7, 2014. Global scientific output doubles every nine years. Blogs.nature.com. http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html. Accessed May 26, 2015.
23. Modha Dharmendra S, Ananthanarayanan Rajagopal, Esser Steven K, et al. Unite neuroscience, supercomputing, and nanotechnology to discover, demonstrate and deliver the brain's core algorithms. *Commun ACM*. 2011;54:62–71.
24. Vendetti Michael S, Bunge, Silvia A. Evolutionary and Developmental Changes in the Lateral Frontoparietal Network: A Little Goes a Long Way for Higher-Level Cognition. *Neuron*. 2014;84:906–917.
25. Lehne Moritz, Philipp Engel, Rohrmeier Menninghaus, et al. Reading a Suspenseful Literary Text Activates Brain Areas Related to Social Cognition and Predictive Inference. *Plos One*. 2015:1–18.
26. Van der Velde, Frank. Where Artificial Intelligence and Neuroscience Meet: The Search for Grounded Architectures of Cognition. Adv Artif Intell. vol. 2010, Article ID 918062.
27. High, Rob. The Era of Cognitive Systems: An Inside Look at IBM Watson and How It Works. REDP-4955-00. December 2012. www.redbooks.ibm.com/abstracts/redp4955.html?open. Accessed September 18, 2015.
28. Su Y, Spangler S, Chen Y. Chemical name extraction based on automatic training data generation and rich feature set. *IEEE/ACM Trans Comput Biol Bioinform*. 2013;10:1218–1233.
29. Su Yan, Spangler Scott, Chen, Ying. Learning to extract chemical names based on random text generation and incomplete dictionary Proceedings of the 2012 ACM SIGKDD conference; 11th International Workshop on Data Mining in Bioinformatics Aug 12-16, 2012. Beijing, China.
30. Lelescu, Anna, Langston, Bryan, Louie Eric, et al. The Strategic IP Insight Platform (SIIP): A Foundation for Discovery. Proceedings of the 2014 Annual SRII Global Conference IEEE Computer Society 2014; 27-34. Washington, DC.
31. Spangler, Scott, Wilkins Angela, D, Bachman Benjamin, J, et. al., Automated Hypothesis Generation Based on Mining Scientific Literature. Proceeding: 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, 2014.
32. Nagarajan, Meena, Wilkins Angela, D, Bachman, Benjamin J. Novikov Ilya, B, Bao, Sheng Hua.

Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW, Australia. August 10-13, 2015.

33. Ban Thomas A. The Role of Serendipity in Drug Discovery. *Dialogues Clin Neurosci*. 2006;8: 335–344.
34. Hargrave-Thomas Emily, Yu Bo, Reyinsson Johannes. Serendipity in Anticancer Drug Discovery. *World J Clin Oncol*. 2012;3:1–6.
35. Riedel Stephan. Edward Jenner and the History of Smallpox and Vaccination. *Proc (Bayl Univ Med Cent)*. 2005;18:21–25.
36. Lohr Steve. And now, from I.B.M., Chef Watson. New York Times. February 27, 2013. http://www.nytimes.com/2013/02/28/technology/ibm-exploring-new-feats-for-watson.html?_r=0. Accessed February 20, 2016.

Address correspondence to: Elenee Argentinis, JD, IBM Watson, 51 Astor Place, New York, NY 10003. E-mail: eargent@us.ibm.com