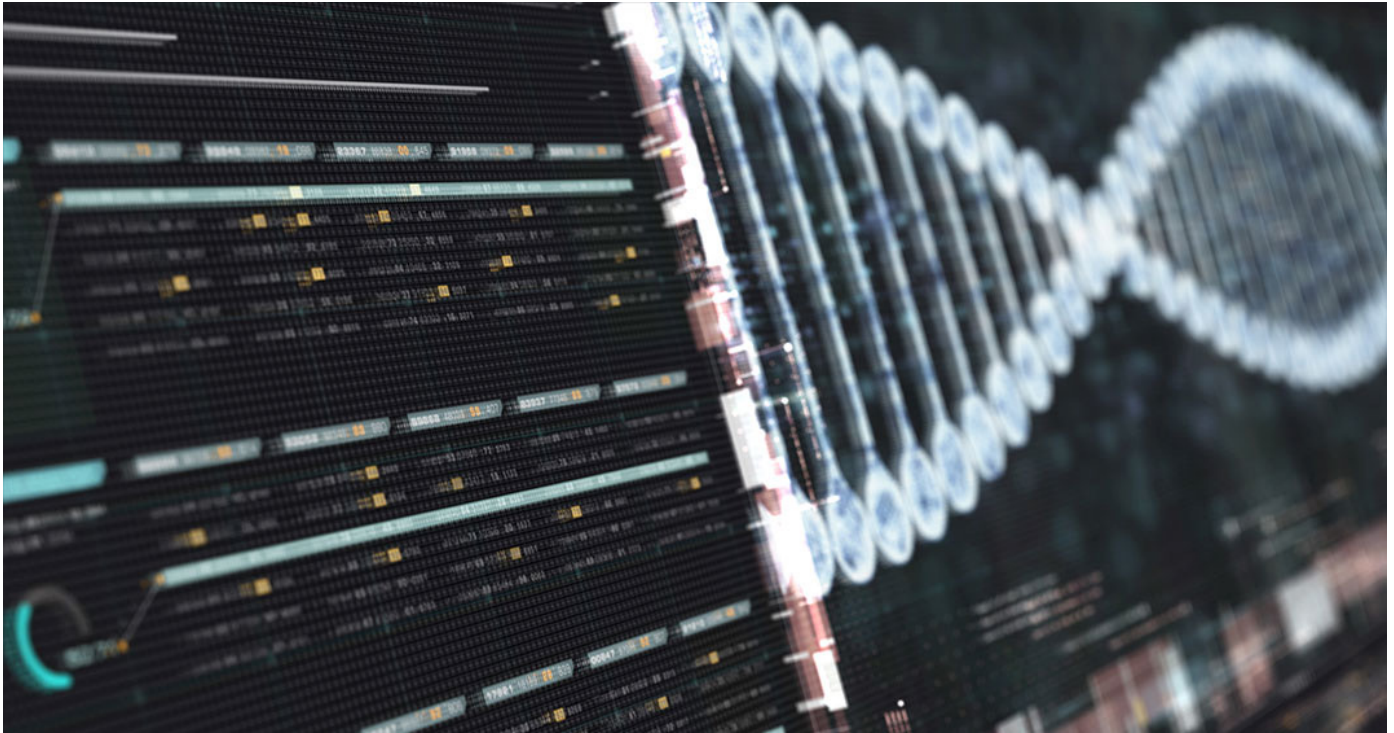# BIG DATA INFECTS LIFE ITSELF

As algorithms reshape the life sciences, who will make it big?
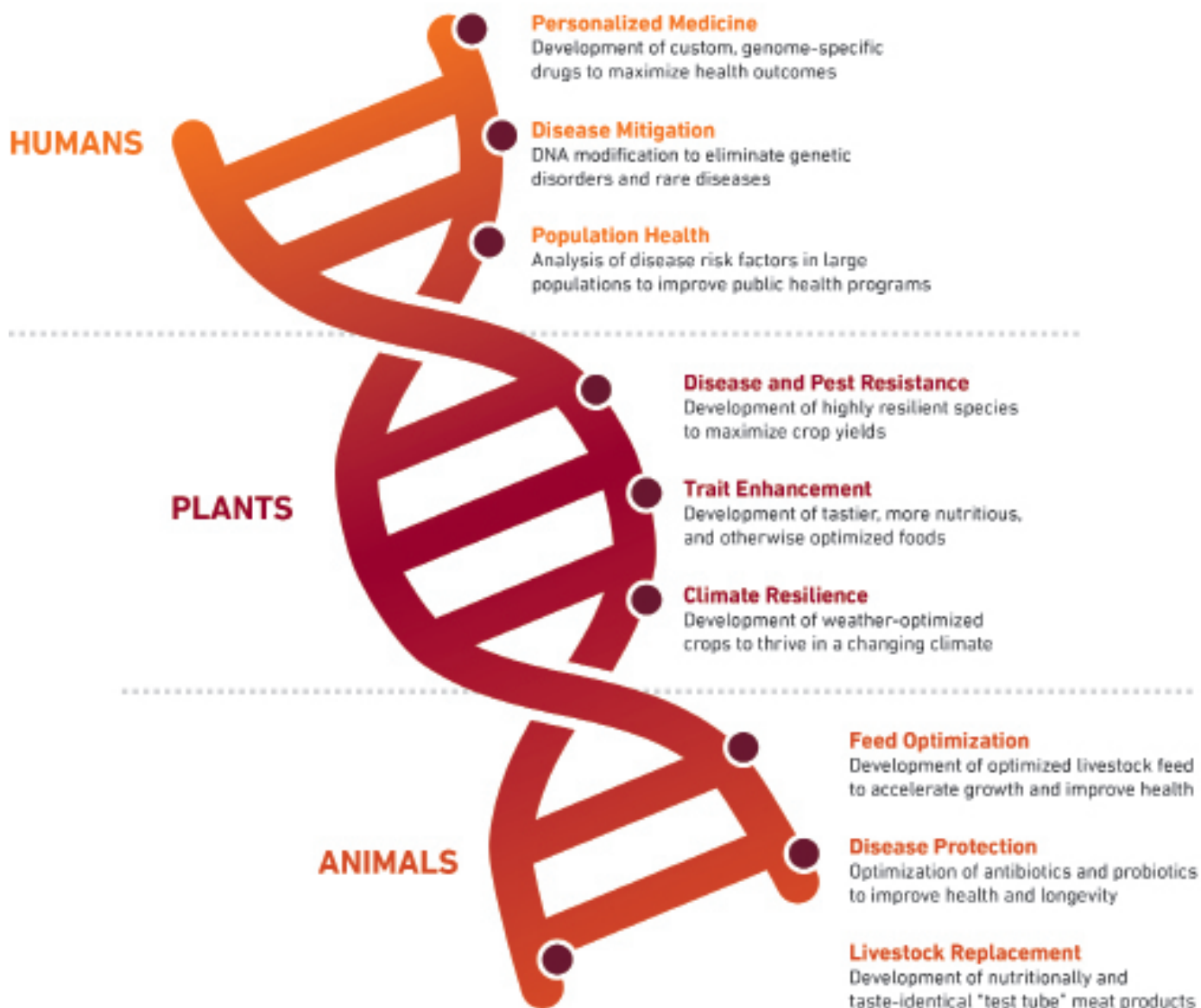
September 2017

■ KEY TAKEAWAYS

- The application of "Big Data" to life science industries such as medicine and agriculture holds tremendous promise, potentially enabling genetic interventions that could dramatically improve human health and agricultural yields.

- While the *collection* of biological data has made tremendous advances in recent years, the *analysis* of that data—a more challenging task—has failed to keep up. However, recent advances in cloud computing, computer processing speed, and artificial intelligence are bringing us near an inflection point where data analytics could "unlock" the true value of biological data.

- This raises the question: can companies currently sitting on large, proprietary biological data sets take advantage of these advances to achieve a Big Data profits explosion like that which internet companies have enjoyed in recent years?

- Life science companies such as Monsanto, Novozymes, Chr. Hansen, and Regeneron could be poised for accelerated revenue growth. In the technology sector, companies such as Dassault Systèmes, Alphabet, and IBM have recently moved into the life sciences, looking to apply their world-class data processing capabilities to biological ventures.

- Given the ever-present chance that biological tinkering will have unintended consequences, as well as the prospect of increased government regulation, investing in this area carries risks.

It is well proven that Big Data can have a big effect on bottom lines. The ballooning values of companies such as Facebook, Alphabet, Tencent, and Amazon.com have shown that if data can be amassed and analyzed at a large scale, it can be monetized for enormous returns—in this case through targeted advertising, social media feed curation, real-time online auctions, and personalized product recommendations. In fact, the collection and analysis of Big Data—that is, massive quantities of information typically of little value in more modest amounts—is considered such a potent competitive advantage that PayPal's CEO Daniel Schulman likened Big Data-loaded algorithms to powerful armaments: "In a world in which algorithms are tremendous weapons, the ammunition for those algorithms is data. And the more data you have, the stronger those algorithms are … It helps you create better value propositions."[1]

Though the profit-generating potential of Big Data has been most visible in the success of internet businesses, tech-savvy firms in most industries have reaped rewards from their investments in data collection and analysis in the form of cost-saving efficiencies, more comprehensive business performance statistics, and more accurate predictive modeling. When it comes to the life sciences—that is, the study of human, plant, and animal biology applied in fields such as medicine and agriculture—advances in the collection and analysis of biological data have the potential to transform society by enabling the development of genetic interventions that could dramatically improve human, plant, and animal health and productivity. They could, in the process, also transform the bottom lines of life science companies that are most successful in these endeavors.

HUMANS

Personalized Medicine
Development of custom, genome-specific drugs to maximize health outcomes

Disease Mitigation
DNA modification to eliminate genetic disorders and rare diseases

Population Health
Analysis of disease risk factors in large populations to improve public health programs

PLANTS

Disease and Pest Resistance
Development of highly resilient species to maximize crop yields

Trait Enhancement
Development of tastier, more nutritious, and otherwise optimized foods

Climate Resilience
Development of weather-optimized crops to thrive in a changing climate

ANIMALS

Feed Optimization
Development of optimized livestock feed to accelerate growth and improve health

Disease Protection
Optimization of antibiotics and probiotics to improve health and longevity

Livestock Replacement
Development of nutritionally and taste-identical "test tube" meat products

The life sciences, however, remain an area where the prospect of large-scale data collection and analysis is particularly mind-boggling due to the sheer magnitude of permutations and combinations involved at the molecular level where biological data is drawn. Consider that the human body has approximately 100 trillion cells, there are 3 billion "base pairs" in the human genome, and many thousands of diseases afflicting humans. In the realm of plant and animal biology, there are approximately 7.8 million animal species and 390,000 plant species, each with their own unique genome.

"What good is collecting all this biological data if we don't know how to use it?"

Despite the inherent technological challenges, there have been significant strides in biological data collection over recent years. The clearest example of these advances is the precipitous decline in the time and cost of human genome sequencing since the first human genome was mapped 14 years ago—a process that, using the technologies available at the time, took some 13 years and cost US$4.9 billion (in 2017 dollars). Today, sequencing a human genome, which is the equivalent of roughly one million pages of printed data, can take as little as 26 hours, while the cost has recently dropped below the highly symbolic US$1,000 mark—long-considered within the genomics field as a benchmark for wide accessibility of the technology. Yet biological data *collection*, such as genome mapping, is just the first stage in the utilization of Big Data in the life sciences, and while there have been some advances in biological data *analysis*—the second stage—computational limitations have thus far prevented researchers from interpreting biological data at the scale necessary to develop the theorized advanced biotechnologies that could revolutionize health care, agriculture, and other life science industries.

■ BREAKING THE BIOINFORMATICS BOTTLENECK

A 2010 paper titled "The $1,000 genome, the $100,000 analysis" by geneticist Elaine Mardis argues that until there are advances in bioinformatics—that is, the use of computer science, statistics, mathematics, and engineering to analyze biological data—commensurate with those in data collection, real-world applications of biological (specifically genomic) data will remain cost-prohibitive.[2] While there have been some promising advances in bioinformatics in recent years, her argument remains largely valid some seven years later. According to Harding Loevner Analyst Chris Mack, CFA, "if there's a bottleneck within the life sciences, it's at the data analysis level right now. Some life science firms have gotten so good at collecting mass quantities of data that they now have a needle in a haystack problem." As Harding Loevner Health Care Analyst Patrick Todd, CFA puts it, "What good is collecting all this biological data if we don't know how to use it?"

"With life science companies today, as with Google and Amazon six years ago—and about every year since—perhaps investors are again underestimating the earnings power of Big Data."

Yet there are signs this bottleneck could soon be relieved. In his recent book *The Gene: An Intimate History*, physician and best-selling author Siddhartha Mukherjee argues that rapid advances in cloud computing and sheer processing speed, combined with breakthroughs in data analysis techniques such as machine learning and other forms of artificial intelligence, could finally enable data analytics that could "unlock" biological data sets. If his thesis is correct, life science firms may be approaching an inflection point where they better understand how to interpret genomic and other biological data, and therefore how to monetize it. Could this development become the next big secular earnings growth theme, doing for life science earnings what Big Data has done for internet earnings?

Harding Loevner Analyst Peter Baughan, CFA, thinks this is entirely possible. "Circa 2010, astute observers would have noticed that ownership of comprehensive data sets, combined with the ability to analyze them efficiently, was of rising strategic value across all industries—a secular beam shining through the cyclical fog," says Baughan. "Still, few could fully anticipate the explosive earnings growth Big Data helped enable. With life science companies today, as with Google and Amazon six years ago—and about every year since—perhaps investors are underestimating the earnings power of Big Data."

■ WHY BIG DATA IS A BIG DEAL

If the projected advances in Big Data analytics do result in rapid earnings growth for life science firms, it will do so via a different path than that travelled by many internet-based businesses. The effective use of Big Data by internet services typically leads to a better user experience that is then converted into higher profits. Amongst internet services that operate on an advertising model, a better user experience should lead to increased web traffic, which inexorably coincides with higher advertising revenues as platform-agnostic advertisers reflexively follow their target demographic across the web. For those that sell services and products such as Uber, Spotify, and Amazon, a better user experience should lead to a larger user base, typically corresponding with increased sales.

For life science businesses, Big Data leads to higher profits primarily via improvements in research and development (R&D) efficiency. This occurs in three ways. First, the use of algorithms and simulations powered by Big Data can enable the gathering of insights at a much faster pace than previous methods—which may include multiple growing seasons or lengthy clinical trials—have traditionally allowed. This, in turn, can enable faster product development. Second, the vastly expanded array of informational inputs resulting from large scale data collection and analysis can lead to more precisely targeted and optimized biological products, increasing their effectiveness and value to the end customer. Third, more accurate predictive modeling made possible by Big Data has the potential to dramatically reduce R&D costs by decreasing the number of failed trials and other "dead ends" inevitably pursued in the course of biological research.
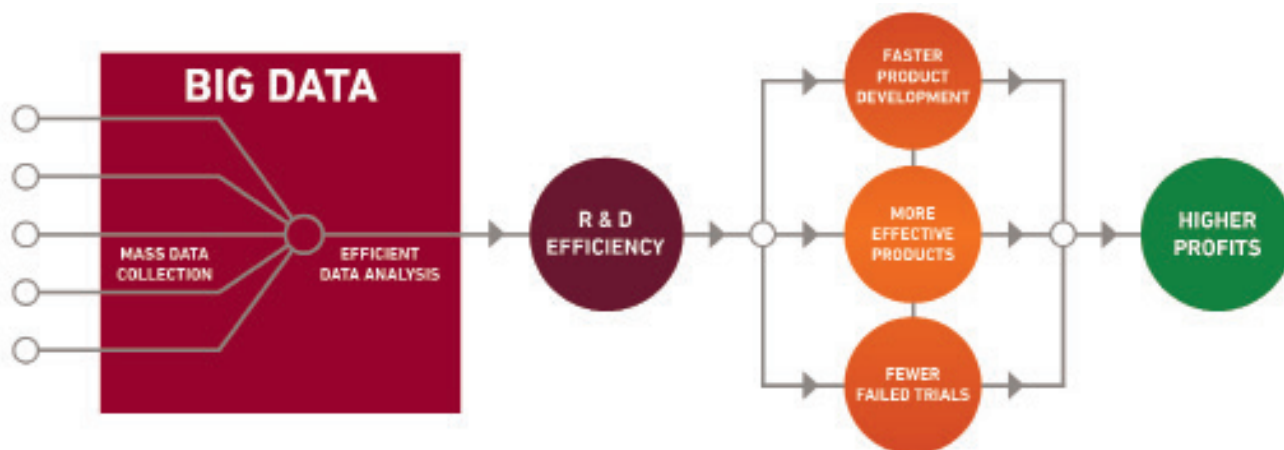
The efficiency of a life science firm's R&D is a critical factor in its success. As pharmaceutical, biotechnology, and other life science companies traditionally spend a high proportion of their earnings on R&D, efficiency in this area can represent a major source of competitive advantage. In short, by bringing more effective and cheaper-to-develop products to market at a faster rate, life science firms that can most successfully exploit the advantages bestowed by Big Data could see significant gains in revenues and market share, especially when combined with the projected high rates of secular growth in life science industries resulting from expanding and aging populations.

Leonard S. Schleifer, CEO of US biotech firm Regeneron, recently affirmed how Big Data can bring about R&D efficiency gains: "PCSK9 is a pretty interesting target; took a lot of smart people like Helen Hobbs and UT Southwestern a long time. We can [now] discover those molecules in five minutes with the database. In five minutes." Schleifer continued, "we view this as a huge competitive advantage. Being able to … either identify [genetic] targets or validate target[s] or tell us not to work on [a] target is enormously valuable. It's still the most important thing in this business, which is coming up with the right targets, and the most powerful way to validate [a] target is genetically."[3]

Big Data provides an additional competitive advantage by way of establishing a formidable barrier to entry. The extraction, storage, cleaning, organization, and interpretation of voluminous data sets is typically both a lengthy and expensive process. This is especially true in biological endeavors, where acquiring and analyzing data requires not only investments in software, data storage, and computing power, but sophisticated lab equipment and specialized technical expertise as well. As Big Data becomes a more critical arena of competition, life science companies with existing biological data sets—especially those that have an ability to analyze them efficiently—should maintain a large lead over new entrants.

■ BIG DATA MEETS SMALL MOLECULE

Life Science Companies

Now that firms can take advantage of the newest advances in data analysis, the question emerges: are there life sciences companies already sitting on large, proprietary data sets that are poised to achieve steep earnings growth similar to those achieved over recent years by internet companies? Monsanto, for example, can now study the precise effects of growing conditions on a large number of potential plant varieties with computer simulations, a task that not long ago required the actual planting of differentiated varieties and waiting an entire growing season before being able to analyze the results. The development of this technology bestows on the agricultural giant a more cost-effective pathway to the development of new plant varieties that are more disease and pest resistant.

Novozymes, a Danish company that controls half the global enzyme market, has a unique "toolbox" of biological screens used to identify, optimize, and mass-produce custom enzymes that can improve industrial processes, livestock and crop yields, and pharmaceutical ingredients. Through a recent partnership, Novozymes now has access to Monsanto's simulation capabilities, allowing Novozymes to analyze the catalyst properties of different enzymes with much greater efficiency.

> **"When you start tinkering with complex, adaptive ecosystems, anything that goes wrong has an almost unlimited potential downside."**

Chr. Hansen, another Danish company, is the global leader in bacteria and cultures with a "library" of over 25,000 in-house strains. The health technology firm currently applies state-of-the-art sequencing and analysis to identify new bacteria strains that can improve human, plant, and animal health. Regeneron is building one of the world's most comprehensive genetics databases, with the goal of sequencing the genetic information of half a million individuals and pairing this data with health records to reveal underlying genetic factors behind health outcomes. The firm believes its ability to use this database to identify the right molecule for drug development—out of millions—and to do so quickly could be a durable competitive advantage.

Technology Companies

Life science companies are not the only firms strengthening their Big Data muscles; major players in the technology industry are beginning to enter the life sciences as well, bringing their vast data-crunching capabilities with them. For example, Dassault Systèmes, the global leader in virtual modeling of complex industrial systems, has recently entered the biological data visualization space, helping its life science customers analyze biological interactions more efficiently.
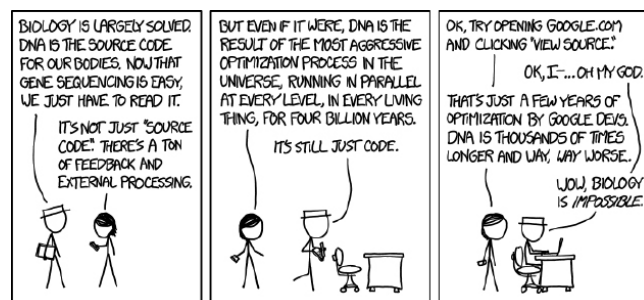
SOURCE: XKCD[4]

Alphabet (Google's parent company) and IBM are also looking to play a major role. Sensing an opportunity to contribute to a range of endeavors through their advanced machine learning and other artificial intelligence technologies, the two technology firms have begun partnering with companies across a variety of industries to help solve complex problems. In the life sciences, Alphabet's Calico division, whose mission is "to understand the biology that controls lifespan," has partnered with C4 Therapeutics, AncestryDNA, AbbVie, and other life science firms to extract insights from biological data that were previously unattainable due to computational limitations. IBM's "Watson Health" program, focusing on genomics, drug discovery, and oncology, takes a similar collaborative approach.

Alphabet is not only helping to crunch other firms' data, it is also collecting its own data through its Verily division, whose goal is to "enable more continuous health data collection for timely decision-making and effective interventions … to radically transform the way healthcare is delivered." In April 2017, Verily launched "Project Baseline," a four-year health data collection exercise with 10,000 participants to identify risk factors for human diseases. According to Harding Loevner Analyst Anix Vyas, CFA, "the real value in biotechnology is somewhere between the raw data itself and the efficient interpretation of that data. A firm need only excel in one of those areas to have a strong value proposition. All the better if it can excel at both." The combination of its world-class data analytics with its proprietary data sets could provide Alphabet with a competitive advantage in the life sciences.

■ BIOHAZARDS

For all its promise, the application of Big Data to life sciences is not without risks. In the development of biological products, one potential hazard is unintended consequences. Vyas explains, "when you start tinkering with complex, adaptive ecosystems, anything that goes wrong has an almost unlimited potential downside. The big danger is if, after considering 1,001 scenarios, a computer tells us that X is the optimal solution, but then scenario 1,002 is what ends up happening in the real world. That's a huge caveat."

A second category of risk is the prospect of increased government regulation. Around the world, regulation is generally highest for biological products designed for humans, followed by animals and then plants, but has been steadily increasing in all categories over previous decades as biological products of all types become more sophisticated. In countries with complex regulatory regimes, bringing new products to market—especially pharmaceuticals and other medical interventions designed for humans—can already take life science companies years due to compliance requirements. This may soon be the case for biological interventions designed for animals and plants as well. As we enter a potential brave new world of genetic manipulation, the probability that governments will impose greater regulatory burdens—perhaps significantly so—on life science businesses is high.

We do not know when the earnings growth of Big Data-oriented life science companies will rise, if at all. Nonetheless we perceive the rate of change in the life sciences to be high and accelerating, and diagnose the value of biological data sets—and the ability to analyze and monetize them—to be greater than the market currently gives credit. If the bioinformatics bottleneck continues to widen, the societal benefits—and the potential growth for life science businesses—could be profound.

■ CONTRIBUTORS

Harding Loevner Analysts Peter Baughan, Chris Mack, Patrick Todd, and Anix Vyas contributed research and viewpoints to this article.

■ END NOTES

[1] Remarks given at the Goldman Sachs "Technology and Internet Conference" on February 15, 2017.

[2] Mardis, Elaine. "The $1,000 Genome, the $100,000 Analysis." Genome Medicine 2, no. 84 (2010): doi: 10.1186/gm205

[3] Regeneron conference call transcript, May 31, 2017.

[4] "DNA" comic available at: https://www.xkcd.com/1605/

■ DISCLOSURES