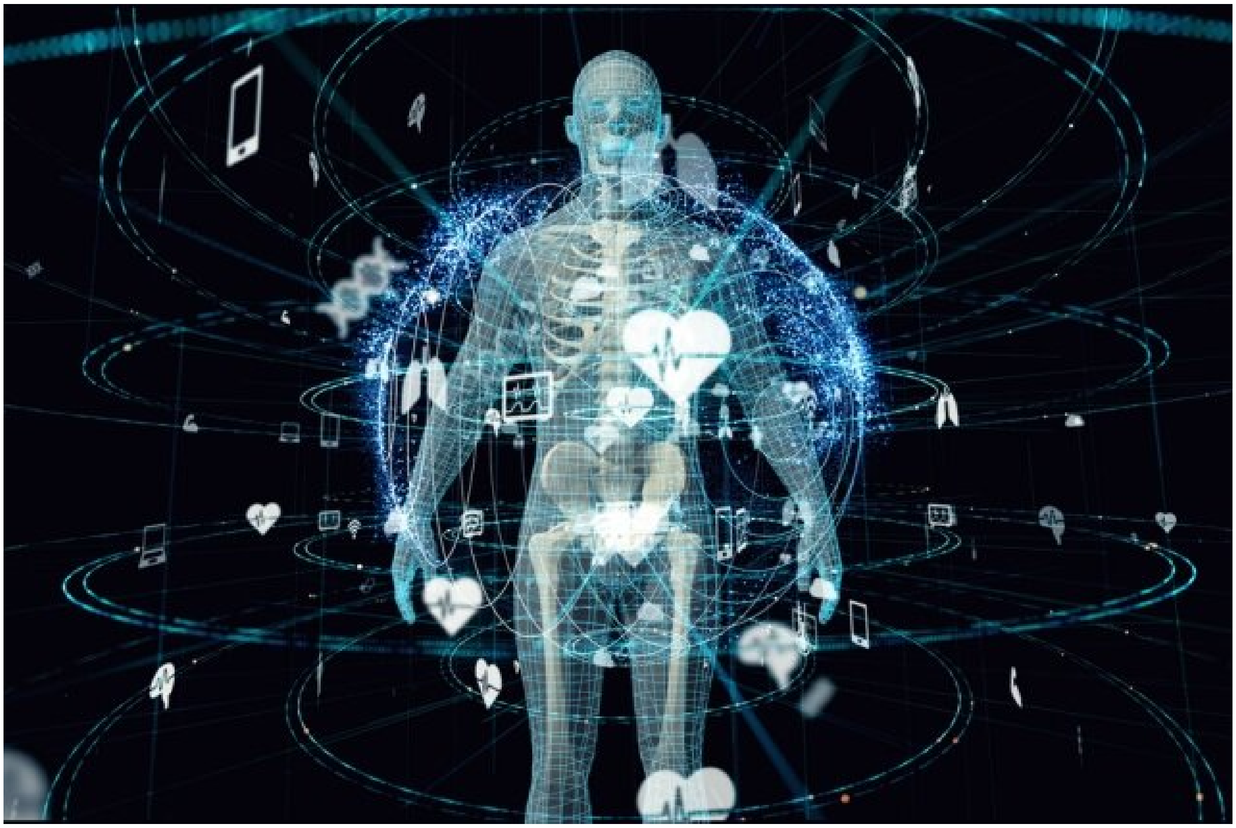
TechXplore

Accelerating life science and health discoveries: Turning data into insights

April 6 2020, by Zach Winn



Paradigm4 allows users to integrate data from sources like genomic sequencing, biometric measurements, environmental factors, and more into their inquiries to enable new discoveries across a range of life science fields. Credit: Massachusetts Institute of Technology

As technologies like single-cell genomic sequencing, enhanced biomedical imaging, and medical "internet of things" devices proliferate, key discoveries about human health are increasingly found within vast troves of complex life science and health data.

But drawing meaningful conclusions from that data is a difficult problem that can involve piecing together different data types and manipulating huge data sets in response to varying scientific inquiries. The problem is as much about computer science as it is about other areas of science. That's where Paradigm4 comes in.

The company, founded by Marilyn Matz SM '80 and Turing Award winner and MIT Professor Michael Stonebraker, helps pharmaceutical companies, research institutes, and biotech companies turn data into insights.

It accomplishes this with a computational database management system that's built from the ground up to host the diverse, multifaceted data at the frontiers of life science research. That includes data from sources like national biobanks, clinical trials, the medical internet of things, human cell atlases, medical images, environmental factors, and multi-omics, a field that includes the study of genomes, microbiomes, metabolomes, and more.

On top of the system's unique architecture, the company has also built data preparation, metadata management, and analytics tools to help users find the important patterns and correlations lurking within all those numbers.

In many instances, customers are exploring data sets the founders say are too large and complex to be represented effectively by traditional database management systems.

"We're keen to enable scientists and data scientists to do things they couldn't do before by making it easier for them to deal with large-scale computation and machine-learning on diverse data," Matz says. "We're helping scientists and bioinformaticists with collaborative, reproducible research to ask and answer hard questions faster."

A new paradigm

Stonebraker has been a pioneer in the field of database management systems for decades. He has started nine companies, and his innovations have set standards for the way modern systems allow people to organize and access large data sets.

Much of Stonebraker's career has focused on relational databases, which organize data into columns and rows. But in the mid 2000s, Stonebraker realized that a lot of data being generated would be better stored not in rows or columns but in multidimensional arrays.

For example, satellites break the Earth's surface into large squares, and GPS systems track a person's movement through those squares over time. That operation involves vertical, horizontal, and time measurements that aren't easily grouped or otherwise manipulated for analysis in relational database systems.

Stonebraker recalls his scientific colleagues complaining that available database management systems were too slow to work with complex scientific datasets in fields like genomics, where researchers study the relationships between population-scale multi-omics data, phenotypic data, and medical records.

"[Relational database systems] scan either horizontally or vertically, but not both," Stonebraker explains. "So you need a system that does both, and that requires a storage manager down at the bottom of the system which is capable of moving both horizontally and vertically through a very big array. That's what Paradigm4 does."

In 2008, Stonebraker began developing a database management system at MIT that stored data in multidimensional arrays. He confirmed the approach offered major efficiency advantages, allowing analytical tools based on linear algebra, including many forms of machine learning and statistical data processing, to be applied to huge datasets in new ways.

Stonebraker decided to spin the project into a company in 2010, when he partnered with Matz, a successful entrepreneur who co-founded Cognex Corporation, a large industrial machine-vision company that went public in 1989. The founders and their team went to work building out key features of the system, including its distributed architecture that allows the system to run on low-cost servers, and its ability to automatically clean and organize data in useful ways for users.

The founders describe their database management system as a computational engine for scientific data, and they've named it SciDB. On top of SciDB, they developed an analytics platform, called the REVEAL discovery engine, based on users' daily research activities and aspirations.

"If you're a scientist or data scientist, Paradigm's REVEAL and SciDB products take care of all the data wrangling and computational 'plumbing and wiring," so you don't have to worry about accessing data, moving data, or setting up parallel distributed computing," Matz says. "Your data is science-ready. Just ask your scientific question and the platform orchestrates all of the data management and computation for you."

SciDB is designed to be used by both scientists and developers, so users can interact with the system through graphical user interfaces or by leveraging statistical and programming languages like R and Python.

"It's been very important to sell solutions, not building blocks," Matz says. "A big part of our success in the life sciences with top pharmas and biotechs and research institutes is bringing them our REVEAL suite of application-specific solutions to problems. We're not handing them an analytical platform that's a set of LEGO blocks; we're giving them solutions that handle the data they deal with daily, and solutions that use their vocabulary and answer the questions they want to work on."

Accelerating discovery

Today Paradigm4's customers include some of the biggest pharmaceutical and biotech companies in the world as well as research labs at the National Institutes of Health, Stanford University, and elsewhere.

Customers can integrate genomic sequencing data, biometric measurements, data on environmental factors, and more into their inquiries to enable new discoveries across a range of life science fields.

Matz says SciDB did 1 billion linear regressions in less than an hour in a recent benchmark, and that it can scale well beyond that, which could speed up discoveries and lower costs for researchers who have traditionally had to extract their data from files and then rely on less efficient cloud-computing-based methods to apply algorithms at scale.

"If researchers can run complex analytics in minutes and that used to take days, that dramatically changes the number of hard questions you can ask and answer," Matz says. "That is a force-multiplier that will transform research daily."

Beyond life sciences, Paradigm4's system holds promise for any industry dealing with multifaceted data, including earth sciences, where Matz says a NASA climatologist is already using the system, and industrial

IoT, where data scientists consider large amounts of diverse data to understand complex manufacturing systems. Matz says the company will focus more on those industries next year.

In the life sciences, however, the founders believe they already have a revolutionary product that's enabling a new world of discoveries. Down the line, they see SciDB and REVEAL contributing to national and worldwide health research that will allow doctors to provide the most informed, personalized care imaginable.

"The query that every doctor wants to run is, when you come into his or her office and display a set of symptoms, the doctor asks, "Who in this national database has genetics that look like mine, symptoms that look like mine, lifestyle exposures that look like mine? And what was their diagnosis? What was their treatment? And what was their morbidity?" Stonebraker explains. "This is cross correlating you with everybody else to do very personalized medicine, and I think this is within our grasp."

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Accelerating life science and health discoveries: Turning data into insights (2020, April 6) retrieved 16 April 2024 from https://techxplore.com/news/2020-04-life-science-health-discoveries-insights.html