

# The data42 program shows Novartis' intent to go big on data and digital

Novartis is betting on both human inquisitiveness and computer expertise to harness the wealth of its digital research and development platform, data42.

By [Goran Mijuk](#) | Mar 11, 2020

Although artificial intelligence has been around for decades, recent developments in the realm of deep learning have enabled data scientists to make surprising leaps. Unlike conventional machine learning algorithms, which learn from the data input, deep learning applies algorithms in layers to create an “artificial neural network” that can learn and make intelligent decisions on its own.

One of the leaders in this emerging field, DeepMind, a unit of the company Alphabet, beat a human player in the highly complex Chinese board game “Go” around three years ago and, using a similar artificial intelligence approach, also trumped biologists in predicting the shape of proteins in 2019.

Such breakthroughs have opened new research avenues and could help solve some of the biggest medical and operational challenges of our time. This includes putting a lid on soaring research and development costs as well as finding innovative ways to treat diseases for which there are only limited or no treatment options available today.

The challenges are indeed big. Bringing a new therapy to market requires investments of more than USD 2 billion today and takes more than 10 years on average. Only 1 out of 10 molecules tested in the clinic reach the market.

Once all the data is curated, the potential to generate new insights is likely to be enormous.

On the medical front, the needs are mounting too. While around 500 drugs were approved in the US in the past decade, medical needs are as high as ever. Many chronic and age-related conditions such as Alzheimer's, for example, remain difficult to treat, and for most of the more than 7 000 known rare diseases, there are no innovative medical options.

“Of course, we don't know yet what we are going to find when we are using this new data and digital technology,” says Pascal Bouquet, Technology Lead at data42. “But we firmly believe we will be able to find insights that are not possible today. We are convinced that we can find nuggets that we have not seen so far and that, in the long run, we can even completely design and discover new drugs based purely on data.”

These hopes have led traditional pharmaceutical players to muscle up their digital expertise and are also attracting new companies such as Google, IBM and Apple into the healthcare space in the hope of developing innovative therapies and disrupting conventional drug development models.

Venture capitalists alone poured more than USD 1 billion into healthcare-oriented artificial intelligence startups in 2018, according to data provider PitchBook. And the market is likely to get hotter. Everest Group, a consultancy, expects overall healthcare investments in artificial intelligence technology to grow from USD 1.5 billion in 2017 to more than USD 6 billion by 2020.

## 2 million patient-years of data

Novartis believes it has an edge in this emerging field. “We have around 2 million patient-years of data in our system,” Bouquet says. “This is the crucial asset which will be instrumental going forward as we apply artificial intelligence tools to sift through the data and find hitherto unknown correlations between drugs and diseases.”

To make this vision come to life, all the clinical and research data – plus potentially real-world data, imaging data and sensor data – first need to be structured and moved to a single platform to create a so-called “data lake.” This is easier said than done, because individual datasets often use different parameters to denote data points such as sex, age, family and disease conditions.

“All of those data need to be cleaned and curated to make them machine-learnable. This is hard and cumbersome work, but it frees up our data scientists to focus on answering questions with data,” says Peter Speyer, who leads product development at data42.

The data size is substantial. The research and development input alone consists of 20 petabytes of data – the equivalent of around 40 000 years of music on an MP3 player.

We have around 2 million patient-years of data in our system. This is the crucial asset which will be instrumental going forward as we apply artificial intelligence tools to sift through the data and find hitherto unknown correlations between drugs and diseases.

## Digging for data nuggets

The team, which involves more than 100 people across the Novartis Institutes for BioMedical Research (NIBR), Global Drug Development (GDD) and Novartis Business Services (NBS) has made great progress so far. They have brought more than 2 000 clinical studies onto the platform and have tested a dozen machine-learning models that could help find new information buried deep in the data.

To gain traction and build proof points, the data42 leadership has set short-term, business-driven objectives that focus on very specific and precise tasks. One such project, which was started recently, aims to identify disease subtypes based on biological characteristics in the area of rheumatoid arthritis.

“For this project, we are working on cleaning the data of our existing trials in this disease domain, which is a task that can be done in a relatively short period of time,” Speyer says. “Our goal is to identify subgroups of high responders to one of our treatments. If we find those, the franchise will potentially be able to set up a new trial and test the findings in the clinic.”



[Discovery](#)

## [Meet Achim, leading the data42 program](#)

[Read More](#)

# Thinking through the question

Among other current projects, the team is also looking at disease progression in certain cancer indications.

And more are yet to come, as the data42 team is working on fine-tuning the data and creating a huge data lake in which to dive for pieces of information that have escaped everyone's attention so far.

"Once all the data is curated, the potential to generate new insights is likely to be enormous," Speyer says. "So, whatever question you have, for example, on heart failure, wherever heart failure is captured as a disease of interest – as a comorbidity or as a side effect – we can pull this into analytics. That is the scalability of data42."

If data42 lives up to its promise, it also has the potential to change how data scientists work together with scientists in the lab and in the clinic. "What you will see is increased collaboration between data scientists, who prepare the data, and medical scientists, who understand the question and what needs to be retrieved from the data," Bouquet explains.

Neither biologists and chemists nor doctors are set to be replaced by the new digital tools, however, which will be only as good as the input they receive. "Sometimes, when you craft a question really well, it turns out the solution is not as complex as you thought," Plueckebaum explained. "You don't need all this artificial intelligence for every question. For some questions you just need to go back to statistics. You find the right data. You apply the right method, and you get the answers. Thinking through the question really helps accelerate and improve the insights – with or without artificial intelligence."

*Main image: Illustration by Philip Buerli*

## AI at Novartis

Artificial Intelligence (AI) at Novartis is a place to innovate, experiment and scale AI-based solutions.

[Learn More](#)

---

**Source URL:** <https://prod1.novartis.com/stories/data42-program-shows-novartis-intent-go-big-data-and-digital>

### List of links present in page

1. <https://prod1.novartis.com/stories/data42-program-shows-novartis-intent-go-big-data-and-digital>
2. <https://prod1.novartis.com/tags/category/discovery>
3. <https://prod1.novartis.com/tags/authors/goran-mijuk>
- 4.
- 5.
- 6.