Modern experimental physics is facing a problem of huge amount of data generated in experiments that needs to be processed in order to draw conclusions from it. One of the biggest centers of experimental physics in the world is CERN, located in Geneva, Switzerland. The LHC (Large Hadron Collider), owned by them, every year produces over 20 petabytes of data. It is used for accelerating particles to the speeds very close to the speed of light and that gives their collisions enough energy to make it possible to encounter some new particles among the products of their decay. For example in order to capture a very elusive Higgs particle scientists have to search through a huge amount of data from many collisions to find just a few ones with Higgs boson as a product.

CERN's current approach to this problem includes using The Worldwide LHC Computing Grid spanning 170 research and computing centers in 35 different countries around the world. This data as available as a live stream to other research centers collaborating with CERN. Most of this data in unstructured, being essentially photographs from light sensors inside LHC, converted into a form which can be analyzed by computer algorithms. First special algorithms written by physicists have to be run. These algorithms enable to sort out only the more interesting events from this mountain of data, the ones that seem the most promising for finding new particles. This enables to shrink the number of data that has to be processed and to focus only on the potentially useful data, separating it from the noise. Of around 300 gigabytes of data per second produced at CERN only around 300 megabytes per second is useful. Then this data has to be processed using various analytical tools. The community of physicists developed their own software toolkit used for this purpose and it is constantly being improved in order to keep up with the continuously growing amount of experimental data. One of these data-processing frameworks is ROOT which is similar to R used it everyday data science.

Physicists are constantly looking for methods which would enable to process even more experimental data and do it as fast as it is possible. Recently they are leaning towards developing better analytical tools as well as cloud computing techniques, hardware improvements in processors and storage and also high performance networks. They have also started to make use of solutions provided by other companies and organizations around the world like cloud services, for example provided by Amazon.

Without these technologies it would be practically impossible to draw any conclusions from experimental data. In order to obtain valuable results and be able to discover new physical phenomena in the future, scientists need to process this data efficiently. As an example, the Higgs boson, experimentally discovered in 2012, confirming the theoretical predictions, would not have been discovered without proper analytical tools developed for dealing with huge amounts of data from LHC. Still there are many physical theories waiting for being confirmed or rejected and in order to understand more deeply the world at the smallest scale of elementary particles, scientists need appropriate tools and techniques provided by research in the field of Big Data and data science.

Cutting-edge data science technologies developed at CERN can also be used by other companies which also struggle with huge amounts of data that needs to be processed. These technologies can also be really helpful in other fields of science, like human genome sequencing project or astronomical observations as well as business, for example applications of Internet of Things or financial analysis.