HPC Innovators - Dr Steven Newhouse, EMBL-EBI

Dr Steven Newhouse is the Head of Technical Services at EMBL's European Bioinformatics Institute (EMBL-EBI).

He has previously held roles in the leadership of European Grid Infrastructure activities and was a Program Manager in the High Performance Computing (HPC) group in the Windows server division at Microsoft. Dr Newhouse's role today involves managing teams that bring both new innovations and established technology services from Europe and around the world to further develop and support the life sciences.

Part of the European Molecular Biology Laboratory, EMBL-EBI helps scientists realise the potential of big data in biology by making the world's public biological data freely available to the scientific community via a range of services, tools, research and training. Dr Newhouse agreed to sit down with the team at Kao Data to discuss his career, where he thinks the future of HPC is headed, and the impact of COVID-19 on data science.

1. How did you get started in your career?

I started off doing my doctoral studies in computational acoustics at Imperial College in London in the 1990s. One of the main problems we ran into at the time was the lack of computing power that was available to us, so my PhD focused on how we could cleverly make use of the limited computing power we had access to. About the same time, high performance computing started to become more readily available, so my postdoctoral work focused on ways to parallelise our code and algorithms so we could solve much bigger problems more effectively using the available HPC machines. A few years later, I moved from being a researcher to being a provider of research computing services, with a focus on how they can be used effectively by scientists, which is the path I am still on today.

2. What are the three most important problems you are trying to solve in your current role?

As Head of Technical Services at EMBL-EBI, I lead the teams that provide a scalable data infrastructure and analysis environment, as well as related services for the data we collect globally. The data management issue is challenging in itself, but the main reason to store the data is to enable further analysis. We have a variety of internal private cloud and computational clusters to help us analyse data and this on-premise infrastructure is increasingly being complimented by public cloud infrastructure. Over the next 10 years this hybrid cloud infrastructure will become commonplace - how we govern and secure our data and our services will be critical.

3. How has HPC evolved to help solve these problems?

The big challenge at EMBL-EBI is the data and not the compute. Many science disciplines are compute intensive, so the more computers they have, the more they can do. Our compute requires us to draw in and process large amounts of data. Our focus is on how we engineer the data infrastructure so that we can get that data quickly to the computers and subsequently write that data back to storage. We've looked at a number of technologies and have been using the dedicated networking common in many HPC machines to improve our storage performance. In addition to high performance computing, we do a lot of high throughput computing, the focus of which is the volume rather than the peak performance of compute. This enables us to balance the bandwidth within our storage network.

4. What does the future of HPC look like five years from now?

We will certainly be at a bigger scale of compute than we are at the moment. Our storage requirements will likely be at the exabyte scale, so how we do data analysis is going to be very interesting. Realistically, it will be a mixture of on-premise and public cloud computing to meet the computing scalability and volume we need. In such an environment there will be some secure, sensitive data we are reluctant to move to cloud platforms. However, the scalability required by the volume of analysis that our researchers will be looking to do will be difficult outside of a hyperscale cloud infrastructure environment.

The challenge then becomes how to make our on-premise storage more accessible to these cloud platforms. Both internal clusters and clusters in the cloud will need to be able to access our storage in a secure, effective and performant manner, regardless of where the work is being conducted. We have experience of solving these problems for our internal environment, but we need to figure out how to scale this offsite.

5. What are you seeing happen in the UK Innovation Corridor that you are most excited about?

The coronavirus pandemic has shined a spotlight on global biological research efforts, which has helped illustrate the role EMBL-EBI plays both in Europe and across the world. We have launched two major new websites recently. Working with the UK Government and UK academics, we launched the **Coronavirus: The Science Explained**. It is a public information site that explains the science of the virus and the pandemic in an accessible way for the interested public.

The second website is the **COVID-19 Data Portal**, which brings together a whole range of COVID-19 related data coming into EMBL-EBI from all over the world. This website pulls in information specific to COVID-19 from different data resources around Europe and the world, organises and presents the data via the portal. The portal also features various tools that can assist with COVID-19 related research.

This is probably the first global pandemic to be so heavily data orientated. The genome of the novel coronavirus was sequenced in a matter of days, demonstrating the speed at which global scientists are operating in relation to the pandemic. Genomics has enabled us to identify at least three different strains and how these strains impact different people is the subject of ongoing research taking place globally as more data is collected on the genetic profile of the individual. One of the big long-term goals of personal healthcare is understanding how an individual reacts to a disease and to certain treatments vs the impact on the general population. We are not there for COVID-19 yet, but the infrastructure for doing that is a lot more capable than it was 5-10 years ago.

More than ever before, the COVID-19 pandemic has demonstrated the power of data sharing, open data and open science, something deeply embedded in the mission of EMBL-EBI. It drives home the importance of data in this era and the need for it to be unconstrained so that science can be conducted to benefit all mankind.

For more information on EMBL-EBI, head to: ebi.ac.uk