Generic components of the eScience ecosystem

At the high energy and cosmic frontiers of subatomic physics ("HEP"), the challenges facing us in dealing with measured and reference data sets are two regular twins: size as well as complexity. By combining the quantitative statistics methods of HEP with generic pattern-matching algorithms, machine learning, and the advances in vectorisation that benefit both GPU and massive multi-processor programming, we stand a chance of dealing with the next generation experiments – if we manage the complex task of integrating these generic components in the large existing scientific code base: as much of a challenge as developing the (very promising) methods and tools themselves.

Yet the speed-up and discriminatory capabilities shown by the new generation of data science tools, the massive multi-core architectures, and the new generation of compilers are essential to get timely results – in an era where multi-messenger science (combining, say, gravitational wave observations, (radio) astronomy, and neutrino telescopes) is about to deepen our understanding of the universe. Can we make the analysis both fast and detailed enough to meet this challenge? And can we get the data to the right place in time to matter?

The latter challenge, getting to the data, and getting scientists themselves to the data, is as complex as getting the data to the processor, if not more so. The HEP community pioneered distributed data management at scale, starting with global catalogues and then inventing the distributed "CVMFS" file system mechanism that now serves code from the Auger cosmic-ray observatory to structural biology. As not only code but also need to be accessible to the research workflows, the new 'data lakes' paradigm – and extensions to CVMFS to make it a global data referral system for a science content delivery network – is about to create an open science platform that frees scientists' workflows from knowing the specific location of data. Applied generically also outside the HEP context, it enables the use of distributed data centres hosted both inside the research community and in public clouds.

In this workshop we explore the possibilities of using generic analysis tools tool sets to get better physics results, look at the complexities that we face when integrating these into existing scientific frameworks, and how we can then leverage the new generation data distribution systems to process data at the moment it matters most. The results of the workshop will then inspire the evolution of physics codes for breaking the data-computing frontier!