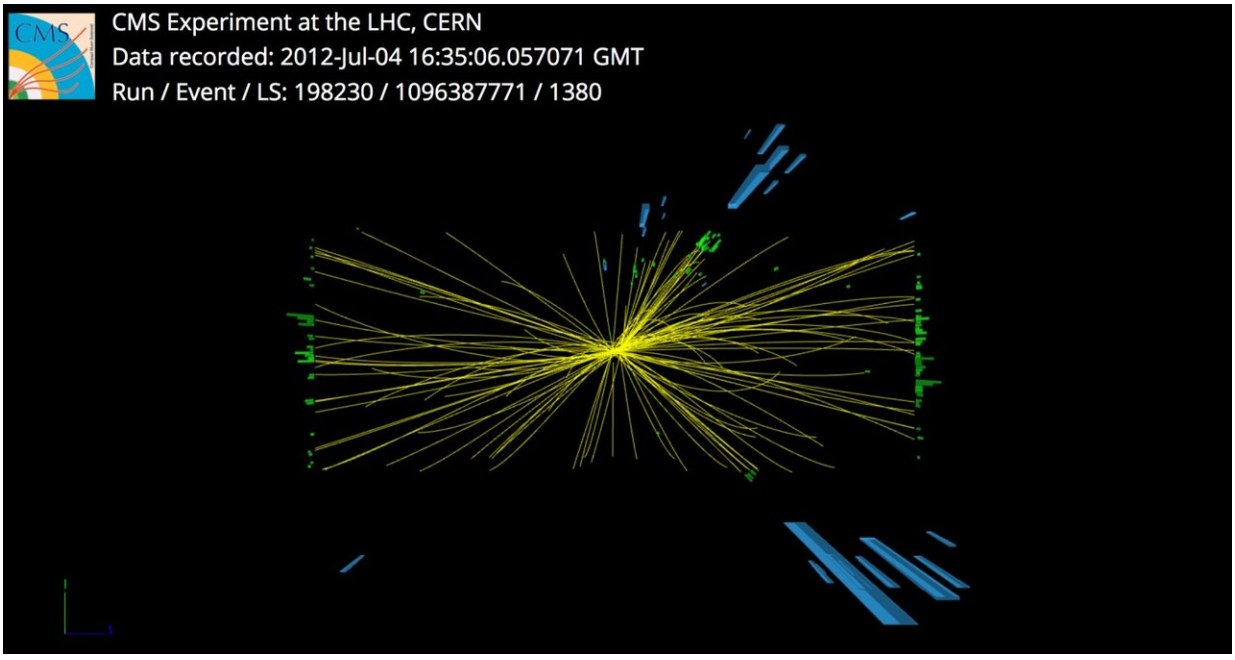
CMS releases more than one petabyte of open data

December 21 2017, by Corinne Pralavorio



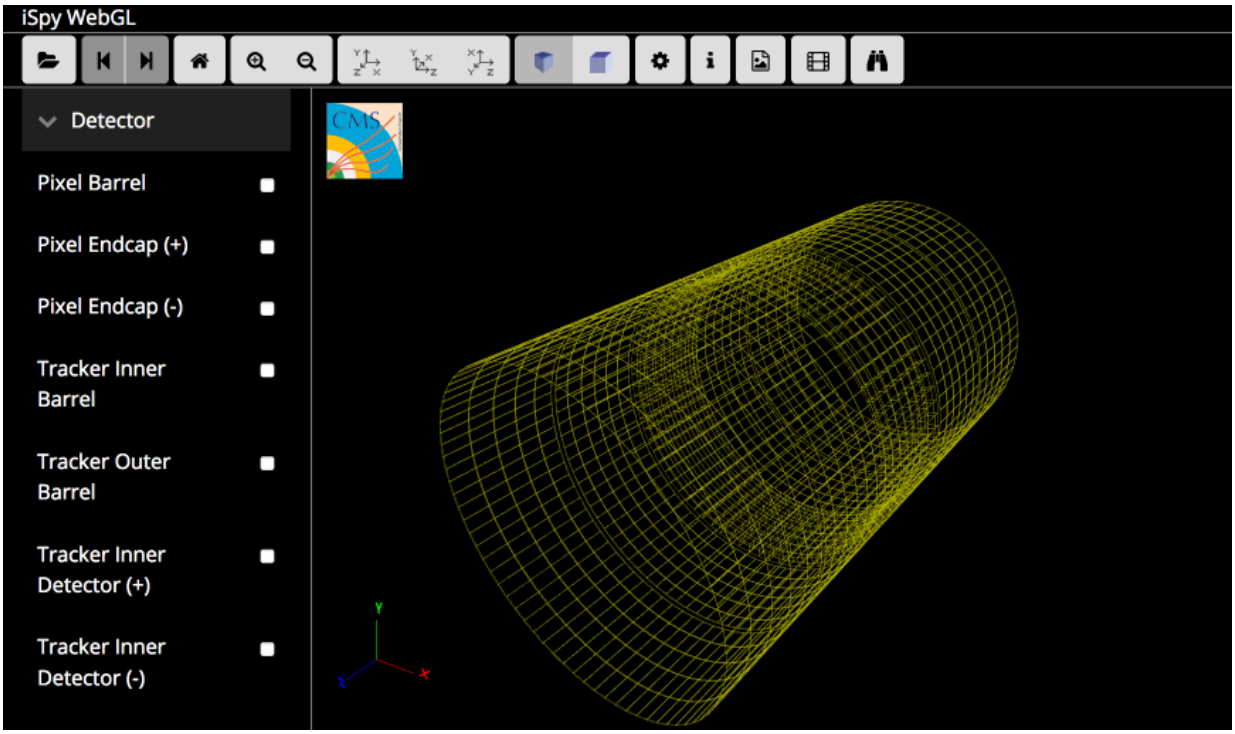
A collision event recorded by CMS in 2012 showing a “Higgs candidate”, available on the CERN Open Data portal with the latest release of CMS Open Data. Credit: Tom McCauley/CMS/CERN

The CMS Collaboration at CERN have just made public around half of the data collected in 2012 by the CMS detector at the Large Hadron Collider. This release includes sets used to discover the Higgs boson, and is being shared through the CERN Open Data portal.

This is the third release of high-level CMS Open Data, following the release of 2010 data in 2014, and 2012 data in 2016. This batch contains more than 550 terabytes of proton-proton collision data recorded at a centre-of-mass energy of 8 TeV as well as around 510 terabytes of Monte Carlo simulation data.

LHC data are complicated and big. CMS researchers have recorded petabytes of data from collisions at the LHC and have so far published hundreds of scientific papers with them. By releasing the data into the public domain, researchers outside the CMS Collaboration have the opportunity to conduct novel research with them.

"Our data are an important element of the CMS Collaboration's rich scientific legacy," says CMS Spokesperson, Joel Butler. "We would like to ensure that they are not only preserved in the long run but are also available to the public, so that both CMS members and external researchers can re-examine them in the future. This is part of our commitment to openness and long-term data preservation."

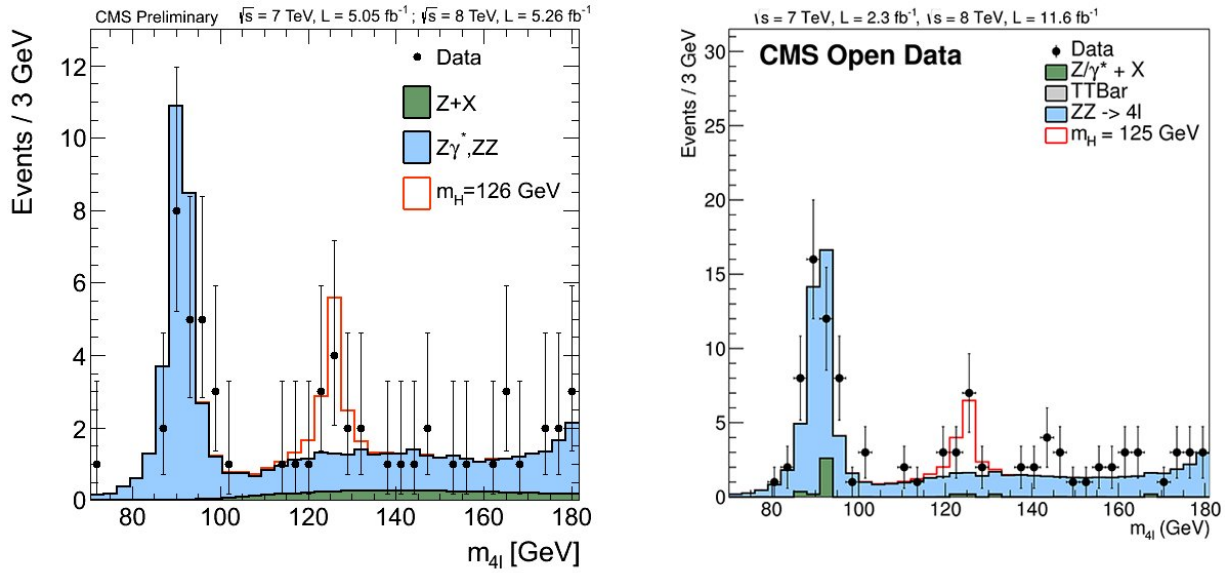


Animation showing a "Higgs candidate" event, recorded by CMS in 2012 and available on the CERN Open Data portal with the latest release of CMS Open Data. Credit: Tom McCauley and Achintya Rao CMS/CERN

Recently, the first two such research papers were published by a team of theorists at MIT interested in performing a measurement CMS scientists had themselves not done: specifically they wanted to measure particular substructures in clusters of particles known as "jets" produced in proton-proton collisions.

The latest release of CMS Open Data also carries the fascinating possibility of allowing people to repeat the analysis that led to the Higgs discovery by studying the same data used by CMS scientists to announce the particle's existence in 2012. As a proof of concept, CMS doctoral student Nur Zulaiha Jomhari analysed CMS Open Data and produced

plots similar to some of those shown when the Higgs discovery was announced. This analysis is a lot less sophisticated than the official CMS one and is not scrutinised by the wider CMS community of experts, but it demonstrates the potential of CMS Open Data.



Left: The official CMS plot for the "Higgs to four leptons" channel, shown on the day of the Higgs discovery announcement. Right: A similar plot produced by Nur Zulaiha Jomhari et al. using CMS Open Data from 2011 and 2012. Although the plots appear similar, the analysis with CMS Open Data uses more data (at 8 TeV and overall) than the official CMS one from the original discovery but is a lot less sophisticated and is not scrutinised by the wider CMS community of experts. Credit: CMS/CERN

In addition to the datasets themselves, the CMS Data Preservation and Open Data team has also assembled a comprehensive collection of supplementary materials, including example code for performing relatively simple analyses, as well as metadata such as information on

how data were selected and what the LHC's running conditions were during the time of data collection.

At the moment, CMS has committed to releasing up to 50% of each year's recorded data a few years after they were collected, once CMS scientists finish most of their analysis of these datasets. "To see our open data in use outside CMS has been very rewarding," says Kati Lassila-Perini, the CMS Data Preservation and Open Access co-coordinator. "It has been a great motivation for us and we look forward to continuing our pioneering efforts to release research-quality open data from the LHC in the years to come."

Provided by CERN