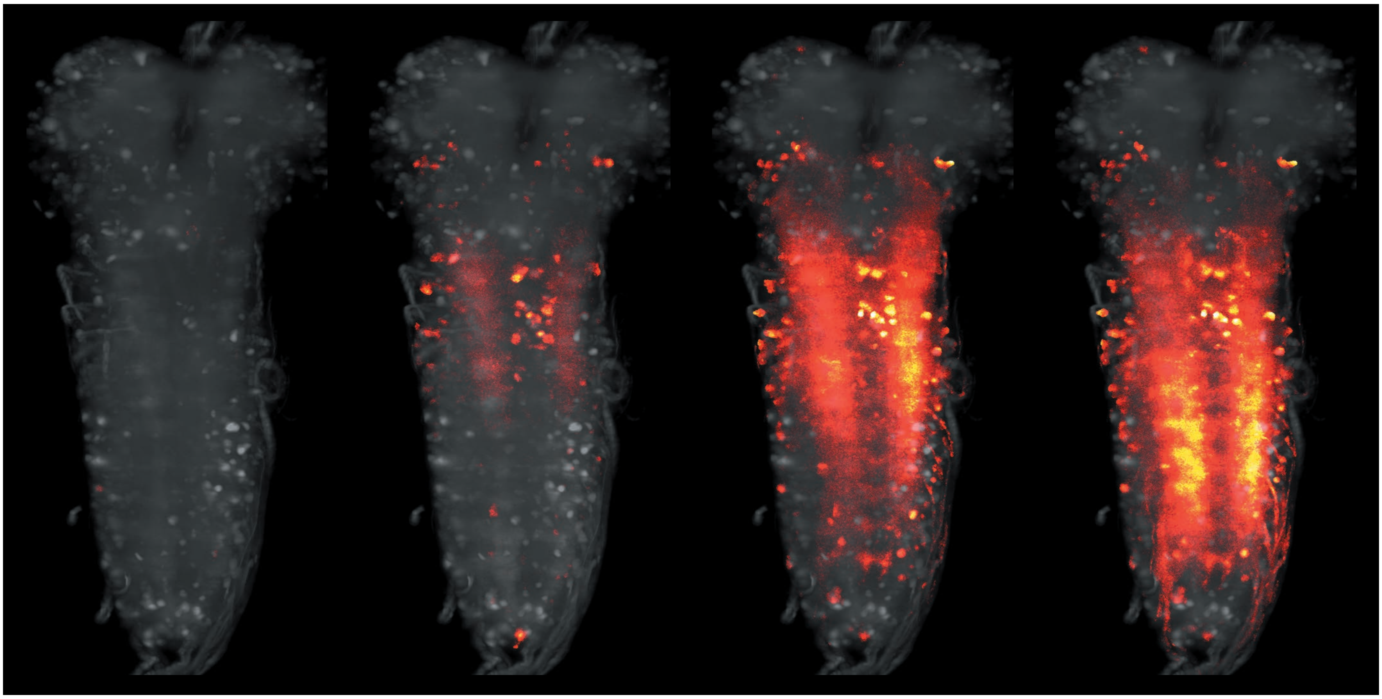
TOOLBOX

THE STRUGGLE WITH IMAGE GLUT

Experiments that generate millions of images have forced scientists to find new ways to store and share terabytes of experimental data.



W.C. LEMON *ET AL. NATURE COMMUN.* **6,** 7924 (2015)

Neurons fire in a fruit-fly larva: a single experiment to track this activity produces millions of images like these.

BY JEFFREY M. PERKEL

As the fruit-fly larva wriggles forwards in the video, a crackle of neural activity shoots up its half-millimetre-long body. When it wriggles backwards, the surge undulates the other way. The 11-second clip, which has been watched more than 100,000 times on YouTube, shows the larva's central nervous system at a resolution that almost captures single neurons. And the experiment that created it produced several million images and terabytes of data.

For developmental biologist Philipp Keller, whose team produced the video at the Howard Hughes Medical Institute's Janelia Research Campus in Ashburn, Virginia, such image-heavy experiments create huge logistical challenges. "We've spent probably about 40% of our time during the past 5 years simply investing in computational methods for data handling," he says. The problem isn't so much storing images — data storage is cheap — but organizing and processing the images so that other scientists can make sense of them and retrieve what they need.

The 'image glut' challenge is becoming an increasing burden for researchers across the biological and physical sciences. Here, Keller and scientists in two other fields — astronomy and structural biology — explain to *Nature* how they are tackling the problem.

MAPPING THE SUN

Somewhere in geosynchronous orbit above Las Cruces in New Mexico, the Solar Dynamics Observatory (SDO) traces a figure-of-eight in the sky. The satellite keeps a constant watch on the Sun, recording its every hiccup and burp with an array of three instruments that photograph the Sun through ten filters, record its ultraviolet output and track its seismic activity. Those data are then beamed to a ground station below. The SDO produces "something like 1.5 terabytes of image data a day", says Jack Ireland, a solar scientist at ADNET Systems, a NASA contractor in Bethesda, Maryland. According to NASA, this amount of data is equivalent to about 500,000 iTunes songs.

To help researchers to stay on top of those images, the ADNET team at NASA, with the European Space Agency, developed the Helioviewer website (helioviewer.org) for browsing SDO images — rather like Google Maps for the Sun, says Ireland — as well as a downloadable application (jhelioviewer.org).

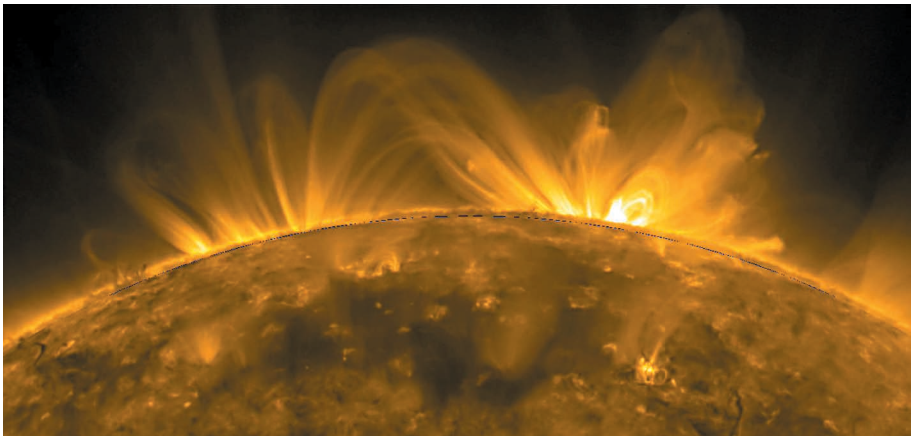
Researchers and astronomy enthusiasts using these tools view not the original data, but instead a lower-resolution representation of them. "We have images of the data," Ireland explains, "not the data itself."

The original SDO scientific images are each 4,096 × 4,096 pixels square and about ▶

12 megabytes (MB) in size. They are taken every 12 seconds, and tens of millions have been collected — a data archive of several petabytes (PB), and growing (1 PB is 1 billion MB, or 1,000 TB). To make images accessible to users, every third image is compressed to 1 MB and made available through Helioviewer.

Users can jump to any particular time since the SDO launched in 2010, select a colour filter and retrieve the data. They can then zoom in, pan around and crop the images, and string them together into movies to visualize solar dynamics. Users create about 1,000 movies a day on average, Ireland says, and since 2011, at least 70,000 have been uploaded to YouTube.

Once they have selected an individual image or cropped area, such as the region around a particular solar flare, users can still download it in its original high resolution. They can also download the complete archive of smaller 1-MB images if they want: but at 60 TB and counting, that process could take weeks.



NASA/SDO

Activity on the Sun seen by NASA's Solar Dynamics Observatory, which gathers 1.5 terabytes of data a day.

FASTER FILE FORMATS

For Keller's developmental-biology group at the Janelia Research Campus, posting their data online for outsiders to access isn't such a concern. If others request it, the team can share images using specialist file-transfer tools, or simply by shipping hard drives. First, however, the team must manage and sort through images that stream off the lab's microscopes at the rate of a gigabyte each second. "It's a huge challenge," Keller says.

Keller's lab uses microscopes that fire sheets of light into the brains and embryos of small organisms such as fruit flies, zebrafish and mice. These have been genetically modified so that their cells fluoresce in response — allowing the team to image and track each cell in 3D for hours. To store its data, the lab has spent around US$140,000 on file servers that provide about 1 PB of storage.

The highly structured organization of the millions of images on those servers keeps the team sane. Each microscope stores its data in its own directory; files are arrayed in a tree that describes the date a given experiment was done, what model organism was used, its developmental stage, the fluorescently tagged protein used to visualize the cells, and the time that each frame was taken. The lab's custom data-processing pipeline was constructed to act on that organization, Keller says.

Yet the directories don't contain the JPEG image files with which most microscopists are familiar. The JPEG format compresses image file sizes, making them easier to process and transfer, but it is relatively slow at reading and writing those data to disk, and is inefficient for 3D data. Keller's microscopes collect images so fast that he needed a file format that could compress images as efficiently as JPEG, but that could be written and read much faster. And because the lab often works on isolated subsets of the data, Keller needed a simple way to extract specific spatial locations or time points.

Enter the Keller Lab Block (KLB) file format, developed by Keller and his team. This chops up image data into chunks ('blocks'), which are compressed in parallel by multiple computer processors[1]. That triples the speed at which files can be read and written, so KLB can compress file sizes just as well as the JPEG format, if not better.

In theory, Keller says, KLB files could be used on commercial digital cameras or on any system that requires rapid data access. KLB source code is freely available, and the lab has made tools and file converters for the MATLAB programming environment and for an open-source image-analysis package called ImageJ, as well as for some commercial packages. Researchers using commercial microscopes could employ the format too, says Keller; he calls it "straightforward" to convert data to KLB files for long-term storage and use.

SHARING RAW DATA

Biologists who take pictures to determine molecular structures also generate vast amounts of image data. And one technique that is growing in popularity — and hence, generating more data — is cryoelectron microscopy (cryoEM).

CryoEM users fire electron beams at a flash-frozen solution of proteins, collect thousands of images and combine these to reconstruct a 3D model of a protein with near-atomic resolution. Most of these reconstructions are less than 10 GB in size, and researchers deposit them in the Electron Microscopy Data Bank (EMDB) — but not the raw data used to create them, which are some two orders of magnitude larger than the resulting models. The EMDB simply was not set up to handle them, says Ardan Patwardhan, who leads the EMDB project for the Protein Data Bank in Europe (PDBe) at the European Bioinformatics Institute (EBI) near Cambridge, UK. As a result, reproducibility suffers, Patwardhan says: without access to raw data, researchers can neither validate others' experiments nor develop new analysis tools.

In October 2014, the PDBe launched a pilot solution: a database of raw cryoEM data called the Electron Microscopy Pilot Image Archive (EMPIAR), also led by Patwardhan. Only data sets for structures deposited in the EMDB are allowed, he says; otherwise, users might be tempted to use the database as a data dump.

EMPIAR currently contains 49 entries averaging 700 GB apiece. The largest is more than 12 TB, and the total collection weighs in at about 34 TB. "We have space available to grow into the petabyte range," Patwardhan says. Users download about 15 TB of data per month in total.

Downloading such large amounts of data presents its own problems: the standard protocol used to transfer files between computers, called FTP, struggles with large data sets; connection loss is common, and download times can slow significantly over long distances. Instead, the EBI has paid for EMPIAR users to access two high-speed file-transfer services, Aspera and Globus Online, both of which transfer data at the rates of "a few terabytes per 24 hours", Patwardhan says. The EBI — which also uses these services to transfer large genomics data sets — pays for its side of the transaction. The cost to the EBI of providing Aspera can be many tens of thousands of dollars per year, he says.

The EMPIAR raw data has already proved its worth. Edward Egelman, a structural biologist at the University of Virginia in Charlottesville, co-authored a study[2] of the structure of an aggregated, filament-like protein called MAVS — which was at odds with another, earlier model of the protein[3]. Egelman proved the earlier structure was incorrect by downloading and reprocessing the raw data set[4]. EMPIAR's grant runs out in 2017, but Patwardhan says that cryoEM researchers have told him they already consider EMPIAR a necessity, and want 'pilot' taken out of the archive's name. "They feel that this should be considered a vital archive for the community — which is nice to hear," he says. ■

1. Amat, F. *et al. Nature Protoc.* **10,** 1679–1696 (2015).
2. Wu, B. *et al. Mol. Cell.* **55,** 511–523 (2014).
3. Xu, H. *et al. eLife* **3,** e01489 (2014).
4. Egelman, E. H. *eLife* **3,** e04969 (2014).