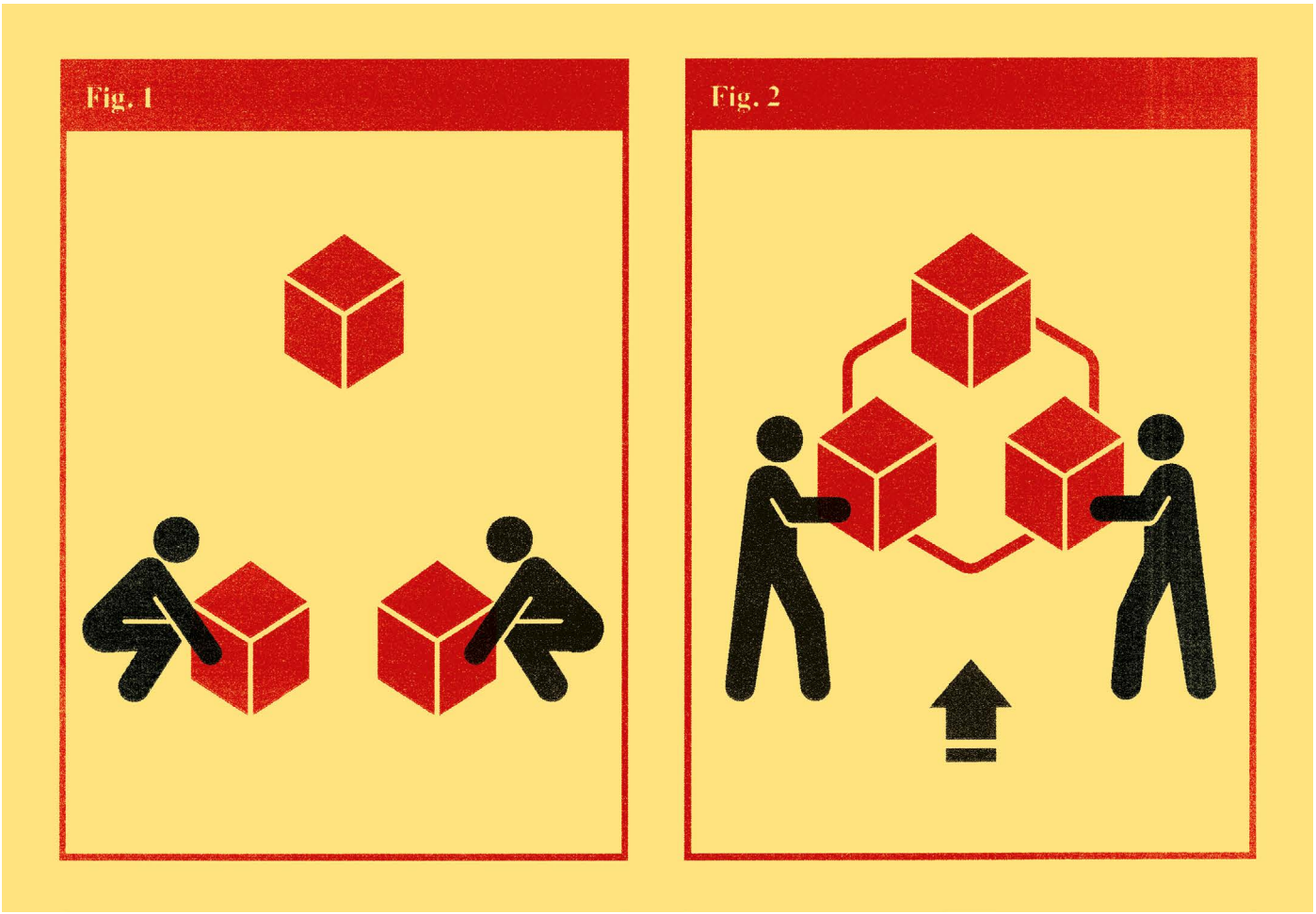

ILLUSTRATION BY THE PROJECT TWINS

ELEVEN TIPS FOR WORKING WITH LARGE DATA SETS

Big data are difficult to handle. These tips and tricks can smooth the way. **By Anna Nowogrodzki**

Big data are everywhere in research, and the data sets are only getting bigger — and more challenging to work with. Unfortunately, says Tracy Teal, it's a kind of labour that's too often left out of scientific training.

"It's a mindset," says Teal, "treating data as a first-class citizen." She should know: Teal was until last month the executive director of The Carpentries, an organization in Oakland, California, that teaches coding and data skills to researchers globally. She says there's a tendency in the research community to dismiss the time and effort needed to manage and share data, and not to regard it as a real part of science. But, she suggests, "we can shift our mindset to valuing that work as a part of the research process", rather than treating it as an afterthought.

Here are 11 tips for making the most of your large data sets.

Cherish your data. "Keep your raw data raw: don't manipulate it without having a copy," says Teal. She recommends storing your data somewhere that creates automatic backups and that other laboratory members can access, while abiding by your institution's rules on consent and data privacy.

Because you won't need to access these data often, says Teal, "you can use storage options where it can cost more money to access the data, but storage costs are low" — for instance, Amazon's Glacier service. You could even store the raw data on duplicate hard drives kept in different locations. Storage costs for large data files can add up, so budget accordingly.

"We can shift our mindset to valuing that work as a part of the research process."

Visualize the information. As data sets get bigger, new wrinkles emerge, says Titus Brown, a bioinformatician at the University of California, Davis. "At each stage, you're going to be encountering new and exciting messed-up behaviour." His advice: "Do a lot of graphs and look for outliers." Last April, one of Brown's students analysed transcriptomes — the full set of RNA molecules produced by a cell or organism — from 678 marine microorganisms such as plankton (L. K. Johnson *et al. GigaScience* **8**, giy158; 2019). When Brown and his student charted

average values for transcript length, coverage and gene content, they noticed that some values were zero – showing where the computational workflow had failed and had to be re-run.

Show your workflow. When particle physicist Peter Elmer helps his 11-year-old son with his mathematics homework, he has to remind him to document his steps. "He just wants to write down the answer," says Elmer, who is executive director of the Institute for Research and Innovation in Software for High Energy Physics at Princeton University in New Jersey. Researchers working with large data sets can benefit from the same advice that Elmer gave his son: "Showing your work is as important as getting to the end."

This means recording your entire data workflow – which version of the data you used, the clean-up and quality-checking steps, and any processing code you ran. Such information is invaluable for documenting and reproducing your methods. Eric Lyons, a computational biologist at the University of Arizona in Tucson, uses the video-capture tool asciinema to record what he types into the command line, but lower-tech solutions can also work. A group of his colleagues, he recalls, took photos of their computer screen's display and posted them on the lab's group on Slack, an instant-messaging platform.

Use version control. Version-control systems allow researchers to understand precisely how a file has changed over time, and who made the changes. But some systems limit the sizes of the files you can use. Harvard Dataverse (which is open to all researchers) and Zenodo can be used for version control of large files, says Alyssa Goodman, an astrophysicist and data-visualization specialist at Harvard University in Cambridge, Massachusetts. Another option is Dat, a free peer-to-peer network for sharing and versioning files of any size. The system maintains a tamper-proof log that records all the operations you perform on your file, says Andrew Osheroff, a core software developer at Dat in Copenhagen. And users can direct the system to archive a copy of each version of a file, says Dat product manager Karissa McKelvey, who is based in Oakland, California. Dat is currently a command-line utility, but "we've been actively revamping", says McKelvey; the team hopes to release a more user-friendly front end later this year.

Record metadata. "Your data are not useful unless people – and 'future you' – know what they are," says Teal. That's the job of metadata, which describe how observations were collected, formatted and organized. Consider which metadata to record before you start collecting, Lyons advises, and store that information alongside the data – either in the software tool used to collect the observations or in a README or another dedicated file. The Open Connectome Project, led by Joshua Vogelstein, a neurostatistician at Johns Hopkins University in Baltimore, Maryland, logs its metadata in a structured plain-text format called JSON. Whatever your strategy, try to think long-term, Lyons says: you might one day want to integrate your data with those of other labs. If you're proactive with your metadata, that integration will be easier down the line.

Automate, automate, automate. Big data sets are too large to comb through manually, so automation is key, says Shoaib Mufti, senior director of data and technology at the Allen Institute for Brain Science in Seattle, Washington. The institute's neuroinformatics team, for instance, uses a template for brain-cell and genetics data that accepts information only in the correct format and type, Mufti says. When it's time to integrate those data into a larger database or collection, data-quality assurance steps are automated using Apache Spark and Apache Hbase, two open-source tools, to validate and repair data in real time. "Our entire suite of software tools to validate and ingest data runs in the cloud, which allows us to easily scale," he says. The Open Connectome Project also provides automated quality assurance, says Vogelstein – this generates visualizations of summary statistics that users can inspect before moving forward with their analyses.

"Our entire suite of software tools to validate and ingest data runs in the cloud, which allows us to easily scale."

Make computing time count. Large data sets require high-performance computing (HPC), and many research institutes now have their own HPC facilities. The US National Science Foundation maintains the national HPC network XSEDE, which includes the cloud-based computing network Jetstream and HPC centres across the country. Researchers can request resource allocations at xsede.org, and create trial accounts at go.nature.com/36ufhgh. Other options include the US-based ACI-REF network, NCI Australia, the Partnership for Advanced Computing in Europe and ELIXIR networks, as well as commercial providers such as Amazon, Google and Microsoft.

But when it comes to computing, time is money. To make the most of his computing time on the GenomeDK and Computerome clusters in Denmark, Guojie Zhang, a genomics researcher at the University of Copenhagen, says his group typically runs small-scale tests before migrating its analyses to the HPC network. Zhang is a member of the Vertebrate Genomes Project, which is seeking to assemble the genomes of about 70,000 vertebrate species. "We need millions or even billions of computing hours," he says.

Capture your environment. To replicate an analysis later, you won't just need the same version of the tool you used, says Benjamin Haibe-Kains, a computational pharmacogenomicist at the Princess Margaret Cancer Centre in Toronto, Canada. You'll also need the same operating system, and all the same software libraries that the tool requires. For this reason, he recommends working in a self-contained computing environment – a Docker container – that can be assembled anywhere. Haibe-Kains and his team use the online platform Code Ocean (which is based on Docker) to capture and share their virtual environments; other options include Binder, Gigantum and Nextjournal. "Ten years from now, you could still run that pipeline exactly the same way if you need to," Haibe-Kains says.

Don't download the data. Downloading and storing large data sets is not practical. Researchers must run analyses remotely, close to where the data are stored, says Brown. Many big-data projects use Jupyter Notebook, which creates documents that combine software code, text and figures. Researchers can 'spin up' such documents on or near the data servers to do remote analyses, explore the data, and more, says Brown. Jupyter Notebook is not particularly accessible to researchers who might be uncomfortable using a command line, Brown says, but there are more user-friendly platforms that can bridge the gap, including Terra and Seven Bridges Genomics.

Start early. Data management is crucial even for young researchers, so start your training early. "People feel like they never have time to invest," Elmer says, but "you save yourself time in the long run". Start with the basics of the command line, plus a programming language such as Python or R, whichever is more important to your field, he says. Lyons concurs: "Step one: get familiar with data from the command line." In November, some of his collaborators who were not fluent in command-line usage had trouble with genomic data because chromosome names didn't match across all their files, Lyons says. "Having some basic command-line skills and programming let me quickly correct the chromosome names."

Get help. Help is available, online and off. Start with the online forum Stack Overflow. Consult your institution's librarians about the skills you need and the resources you have available, Teal advises. And don't discount on-site training, Lyons says: "The Carpentries is a great place to start."

Anna Nowogrodzki is a journalist based near Boston, Massachusetts.