TOOLBOX

# WAYS TO AVOID A DATA-STORAGE DISASTER

*Hard-drive failures are inevitable, but data loss doesn't have to be.*

ILLUSTRATION BY THE PROJECT TWINS

BY JEFFREY M. PERKEL

Tracy Teal was a graduate student when she executed what should have been a routine command in her Unix terminal: **`rm -rf *`**. That command instructs the computer to delete everything in the current directory recursively, including all subdirectories. There was just one problem — she was in the wrong directory.

At the time, Teal was studying computational linguistics as part of a master's degree in biology at the University of California, Los Angeles. She had spent months developing and running simulation software, and was at last ready to begin her analysis. The first step was to "clean up the data and get organized", she says. Instead, she deleted her entire project. And unlike the safety net offered by the Windows and Macintosh operating-system trash cans, there's no way of recovering from executing **`rm`**. Unless you have a backup.

In the digital world, backing up data is essential, whether those data are smartphone selfies or massive genome-sequencing data sets. Storage media are fragile, and they inevitably fail — or are lost, stolen or damaged.

Backup options range from USB memory sticks and cloud-based data-storage services to huge institutional magnetic-tape servers, with researchers typically exploiting more than one. But not all such strategies have the same advantages, and scientists must discover what works best for them on the basis of the nature and volume of their data, storage-resource availability and concerns about data privacy.

In Teal's case, automation saved the day. The server on which she was working was regularly backed up to tape, and the "very friendly and helpful IT folks" on her department's life-sciences computing helpdesk were able to recover her files. But the situation was particularly embarrassing, she says, because Teal — who is now executive director at The Carpentries, a non-profit organization in San Francisco, California, that runs workshops on scientific computing — had previously worked for the information technology ▶

(IT) team. It was "like the lifeguard having to be rescued", she says.

Here are 11 tips that could make potential data-loss disasters a little less painful.

**1. Apply the 3-2-1 rule.** The rule of thumb to follow when making data backups, says Michael Cobb, director of engineering at DriveSavers, a data-recovery firm in Novato, California, is '3-2-1': "It's three copies, [on] two different media, one off-site." You might, for instance, maintain copies on your personal computer, an external hard disk and the cloud-based file-synchronization service Dropbox (US$12.50 per user each month, for 3 or more users and 3 terabytes of storage). "This is a rubric to take inspiration from, not a law," notes Elizabeth Wickes, an information scientist at the University of Illinois at Urbana–Champaign — precious data might require extra precautions.

**2. Talk to the specialists.** Your institution employs people to think about data full-time, so talk to them, advises Juliane Schneider, who leads data curation at Harvard Catalyst in Boston, Massachusetts. Your research-computing centre might offer free or low-cost institutional backup systems; your librarian can help you to craft a data-management strategy; and your grants office can advise you on funding-agency requirements, including how, and for how long, data must be maintained. "They want to help you keep your data — especially if you have a grant," she says.

**3. Manage your data.** Reliable backups require clever data management. Referencing the organizing method devised by Marie Kondo, a popular Japanese lifestyle consultant and author of *The Life-Changing Magic of Tidying* (2014), Ciera Martinez, a data scientist at the University of California, Berkeley, advises asking of each file: 'Does this need to be stored?' Adds Teal, with a laugh, "You can't just keep the data that bring you joy!"

Establish conventions on file naming and organization — for instance, that each project gets its own folder; that data and code go into dedicated subdirectories; and that each project folder includes a file that documents the project's aims, methods, metadata and files. Plan where and how data are backed up, and develop a schedule — daily or weekly, for instance — for doing so.

Raw data should always be saved, but intermediate processing files can often be discarded. Massive data sets require special thought: some cloud-based providers cap the sizes of stored files, and data-transfer and storage costs can become prohibitive.

**4. Safeguard privacy.** Data gathered from patients or students are often restricted, which means that they cannot be stored just anywhere. At her institution, Wickes says, researchers have several cloud-based options for data backup, but only one is approved for use with sensitive data. Your department's IT team can offer advice. "Being out of compliance for data protection can be very serious. You could face financial penalties, or lose the ability to conduct research," Wickes says.

**5. Automate backing up.** When making backups, automation is key. Kelly Smith, a cardiac geneticist at the University of Queensland in Brisbane, Australia, has access to a shared network drive that is copied to tape. She used to move her files to the drive manually, but only monthly; in the event that the drive failed, newer files could be lost. An automated cloud-based backup system called Druva inSync, from data-protection firm Druva in Sunnyvale, California, now obviates that concern. "It's one less thing I have to worry about," she says.

"You have to not think about it," explains Teal. "Because when you're most stressed is when things go down, and when you've forgotten backups for the past three months."

**6. Protect raw data.** All data are precious, but raw data are irreplaceable: the only way to recreate them is to run the experiment again. These must therefore be backed up — and kept as read-only files. Wickes once had to kill a project because she opened a crucial file in Microsoft Excel, which automatically formatted a column, changing the values and ruining the underlying data set. So, protect your raw data, says Martinez, "no matter what".

**7. Make backing up achievable.** A data-management plan must be easy to follow for new members of the lab, as well as for postdocs who are pulling an all-nighter. "You might say, 'Oh, this is a perfect system.' Okay, now, are you going to do it at 3 a.m., after you've been working for 24 hours on something? Are you going to do it when you're in the middle of fighting with a code problem?" Wickes says. Discuss the strategy as a team, and make sure that it's workable. Then, just as you would for your −80 °C freezer, simulate what would happen if disaster strikes: what data would you lose, and how quickly could you recover? "At a minimum, that as a thought experiment would be valuable," says Teal.

**8. Test backups regularly.** Don't assume that your backups are working: test them. Can you open your files? Do you have the necessary applications, login credentials and registration keys to run them? Wickes' departmental IT service offers staff a free account on CrashPlan from Code42 Software in Minneapolis, Minnesota, which automates backups to the cloud. One day, Wickes decided to test her backup, only to find that it had stopped syncing six months earlier. "I was fine because I had a local Time Machine backup, as well," she says, referring to Apple's backup system for computers running its Macintosh operating system. Reiterating the advice he gave in tip 1, Cobb says: "So, 3-2-1 backup, and then restore [some key files]. And test it on a different computer, in a different room, on a different device — because if the worst-case scenario happens, you won't have your device."

**9. Expect the unexpected.** Life happens. Cobb — who lost all of his personal possessions in a wildfire in 2017 — had a client who stored a rack of 96 hard disks underneath a fire-control sprinkler. One day, the sprinkler popped, and the disks were inundated with water. "None of that data was backed up," he says. Leslie Vosshall, a neurobiologist at the Rockefeller University in New York City, almost lost her mosquito-genome sequencing data in 2012 when her basement servers were flooded in the wake of Hurricane Sandy. Such events are unavoidable but can often be anticipated — so search hard for vulnerabilities. About a year-and-a-half ago, Cobb's office was shaken by a small earthquake — hardly a surprise in California. A picture of former US president, and one-time client, Gerald Ford fell off the wall and hit his laptop "just right", shattering the screen. "After that, I was like, 'I better move things around so I am better prepared'."

**10. Keep a backup offline.** Internet-connected backup devices are convenient: the data are instantly available. But those devices are also instantly vulnerable to user error and malicious software (malware). Craig Rager, chief technology officer at Data Mechanix, a data-recovery firm in Irvine, California, says that many of his clients have suffered ransomware attacks, in which a virus encrypts a computer's hard disk, making it unusable. A backup drive, whether attached to the computer directly or through a network, can also be hit in such an attack, he notes. "Because you can never eliminate this threat 100%, the only thing you can really do is have a device that you back up, which is then taken offline or not accessible to your network," for instance, by being powered off.

**11. Plan ahead.** Ultimately, your data need to be available in the future. So, think about "future you", says Teal. Consider the media on which your data are saved, and the applications that you use to open them. Try to stay up to date. Much of Vosshall's early data are stored in an obsolete disk format, she says, meaning they're backed up but inaccessible. "I'd have to go to an antique store to find a reader." Even the cloud provides no guarantees: data-storage companies can shift their business priorities, or you might simply lose access to your account. So, make sure to keep a local backup — or, at least, back up your data on independent services. "People will ask, 'You mean, you don't trust Google Docs?'" says Wickes. "It's not necessarily about trusting Google Docs, it's about trusting that you don't lose access." ■

**Jeffrey M. Perkel** *is technology editor,* Nature.