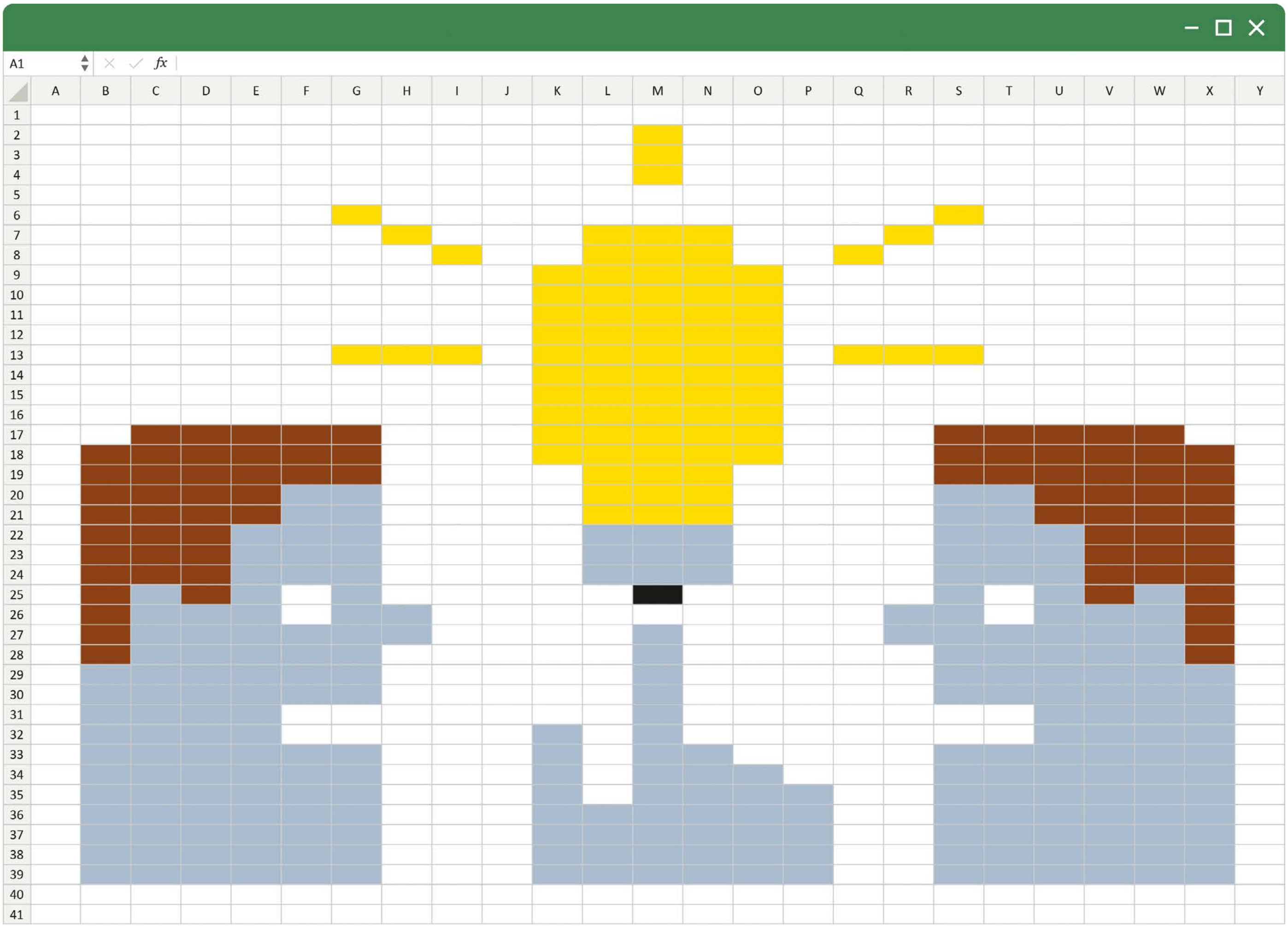

ILLUSTRATION BY THE PROJECT TWINS

SIX TIPS FOR BETTER SPREADSHEETS

Microsoft Excel and Google Sheets are powerful and widely used. But there's a right way and a wrong way to use them, data scientists say. **By Jeffrey M. Perkel**

As a data-science librarian at the University of California, San Diego, Stephanie Labou has seen her share of spreadsheet horror stories. The most haunting was a table of hand-entered GPS coordinates.

"It was a complete mix," Labou recalls. The spreadsheet was produced by citizen-scientists. Some had written the word 'degrees', some '0' and some 'o'; some had used superscripts, some hadn't; others wrote 'north', 'west' or neither. "We're talking like tens of thousands of rows of data, where every single latitude and longitude was annotated differently," she says. "That was the least consistent spreadsheet I've ever seen."

Data scientists express strong feelings about using spreadsheets for data analysis. On the whole, they prefer programming languages such as R and Python, in which analyses are more easily documented and more reproducible. But many researchers are more comfortable with spreadsheets, and being shamed for using them is counterproductive, says Labou. Sometimes, spreadsheets are the fastest way to solve a problem. And there is really no other option for recording tabular data.

Spreadsheets are reactive: cells that depend on other cells will update automatically as the data change. They can also be helpful, intelligently formatting data to make them easier to read. Plus, they are everywhere. Spreadsheets are "where data science begins", says Tracy Teal, open source program director at the software developer RStudio in Boston, Massachusetts.

But they are also trickier than they seem. A function to take the average of a column, for instance, will return the wrong value if the formula fails to account for the correct data range. Cells that seem empty might not be. And autoformatting doesn't always work as expected. Researchers have long known that some genomic studies contain garbled data because Excel improperly converted some gene symbols, such as *OCT4*, into dates. An analysis of around 11,100 papers published between 2014 and 2020 found that 31% still include such errors (M. Abeysooriya *et al. PLoS Comput. Biol.* **17**, e1008984; 2021).

As data scientists Karl Broman at the University of Wisconsin–Madison and Kara Woo, then at the University of Washington, Seattle, wrote in 2018: "Spreadsheets, for all of their mundane rectangularness, have been the subject of angst and controversy for decades" (K. W. Broman and K. H. Woo *Am. Stat.* **72**, 2–10; 2018)

Here are six tips for using them correctly.

Keep raw data raw

Christie Bahlai, a computational ecologist at Kent State University in Ohio, has helped to create workshops and teaches courses on best spreadsheet practices for ecologists. She says her number-one piece of advice is to "keep your raw data raw".

Spreadsheets, Bahlai says, are "tactile": they are user-friendly, intuitive and easily manipulated. But they are also "easy to mess up", and it is "easy to lose track of what you've done". An errant mouse click can cause data to end up in the wrong place. And the autoformatting function can ruin the data. Furthermore, the spreadsheet can contain organization information that might not be immediately clear. As a result, Bahlai recommends that users make their original spreadsheet a read-only document and work on copies, so that they can start over if necessary.

Bahlai recalls one case in which she kept finding single letters in one of the spreadsheet's columns as she began to process the data. "I'm like, 'what does 'M' mean? What does 'A' mean?'" It turns out that a team member had typed 'NO SAMPLE' vertically in one of the columns, one letter per row — an organization decision that is clear to a human reader, but not a computer. When she sorted the table, that visual organization was lost. "It was like solving a jumble," she says with a laugh. "I realized, 'Oh, this spells something, there's a message!'"

Make data machine-readable

Spreadsheets provide extensive formatting options, from font styling to background fills to borders. This digital 'bling' can liven up a table and make it more readable. But when researchers use such styling to encode data, they can run into trouble.

"My top piece [of advice] is, do not encode data with colour or formatting, create another column that can be sorted or filtered," says Mine Çetinkaya-Rundel, a statistician at Duke University in Durham, North Carolina.

That is because cell formatting is difficult for downstream users to capture. "All the tools available to data scientists are unaware of data expressed as formatting rather than as text or numeric values," says Duncan Garmonsway, a data scientist in the UK Government Digital Service in Lincoln. Formatting can be lost during routine table manipulations. And researchers might struggle to remember what the formatting represents when they return to the spreadsheet months or years later.

Luis Verde Arregoitia, a mammalogist at the Institute of Ecology (INECOL) in Veracruz, Mexico, experienced that when he revisited an old collection of biodiversity records. He had highlighted rows in yellow, orange or green to indicate his level of trust in the data. "At this point," he says, "I don't really remember the exact colour-coding scheme that I was using."

Be consistent

Data-analysis tools expect spreadsheets to be in a specific format: one row of column titles, no merged cells and one table per page. Ideally, all cells are filled, even when there are no data (for instance, with 'NA'), and contain precisely one piece of data. To tabulate data from a field study to count insects, for instance, use separate columns for insect types and for the count, says Teal, instead of, say, '3 red beetles'.

Specialized tools can untangle spreadsheets that deviate from the ideal. Verde Arregoitia's 'unheadr' package, for instance, handles tables that include rows to subdivide a table into different groups, which he calls 'embedded subheaders'. Garmonsway's 'tidyxl' and RStudio data scientist Jenny Bryan's 'googlesheets4' provide ways to extract the formatting.

"Writing a roadmap for yourself is important."

The most important thing, Labou says, is consistency — decide on an approach, document it and stick to it. How will species be indicated? And how should dates be formatted — does '2/1/2022' mean 1 February or 2 January? Most experts recommend either the YYYY-MM-DD format — the International Organization for Standardization standard — or dedicating separate columns to year, month and day. When combined with data validation, the use of separate columns means "there's absolutely no ambiguity", Labou says. But, warns Broman, it does make it more difficult to compute date differences.

Document your work

Whereas programming scripts can be saved and version-controlled, keystrokes and mouse-clicks generally cannot be. But spreadsheet users can still document their analyses.

Designate a spreadsheet (or tab) as a 'code book' that documents abbreviations, how data were collected, units of measurement, how missing values will be represented, the calculations being performed and any metadata needed to understand, process and maintain the spreadsheet. "Writing a roadmap for yourself is important," says Çetinkaya-Rundel.

Then, says Bahlai, "write the recipe of what you've done to your data". What does each formula do, and where does it draw its data from? "You will regret it if, when you go to write your methods and you go, 'Huh, how did I take the average of this?'" she says. (In Excel, you can use the 'audit' function to see the flow of data through the formulae, notes Felienne Hermans, a computer scientist at Leiden University in the Netherlands.)

Cross-check your data

Data analysts often add cross-checks to ensure that their data-processing code works as expected. Spreadsheet users can do something similar, says Hermans.

In a study with samples from both cases and controls, for instance, the total number of values in the two groups should always equal the number of samples; if nothing else, that cross-check ensures that cells that you think are empty actually are. "Building in some of these cross-checks so you can see that everything is in order, that's actually a really, really good idea," she says.

You can also 'protect' parts of the spreadsheet from modification, and apply data validation to ensure that date columns contain valid dates, that numbers fall within certain ranges or that text fields include expected terms. Alternatively, suggests Çetinkaya-Rundel, use a data-entry form (such as a Google Form) rather than editing the spreadsheet directly. That way, values can be checked as they are entered, and users cannot accidentally modify the document. Finally, says Teal, double-check your work. Data analysis is often iterative, she notes. "You don't just walk in the door and go, 'I am going to do this equation,' sit down, do it, done." So, once you've settled on a workflow, reset and start over, she says, and just make sure that you have the answer that you thought you did.

Think ahead

The good news is, data scientists can generally wrangle spreadsheets irrespective of their format. "A key principle that I have as a data analyst is, if someone asks me in what form would I like the data, I always say 'in their present form'," says Broman. "If the data need to be reorganized or transformed in some way, I'm always in the best position to do that." But it's better, Labou says, to work out what you hope to do with your data before creating your spreadsheet in the first place. Which variables and covariates will you be using? What time steps do you need? What analyses will you be performing? "Thinking that through ahead of time, is one of the best things that people can do," she says.

And consult your collaborators, Garmonsway adds. Rules for data organization "aren't carved in stone anywhere", he says. "Physicists didn't discover them in the fundamental laws of the Universe. They emerged because it's hard to work with other people. So if you collaborate when you create your spreadsheet, it's much more likely to be useful to other people, because it's already useful to someone who isn't you."

Jeffrey M. Perkel is the technology editor at *Nature*.