Data carpentry

WRITTEN BY DAVID MIMNO

The New York Times has an article titled For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. Mostly I really like it. The fact that raw data is rarely usable for analysis without significant work is a point I try hard to make with my students. I told them "do not underestimate the difficulty of data preparation". When they turned in their projects, many of them reported that they had underestimated the difficulty of data preparation. Recognizing this as a hard problem is great.

What I'm less thrilled about is calling this "janitor work". For one thing, it's not particularly respectful of custodians, whose work I really appreciate. But it also mischaracterizes what this type of work is about. I'd like to propose a different analogy that I think fits a lot better: *data carpentry*.

Note: data carpentry seems to already be a thing

Why is woodworking a better analogy? The article uses a few other terms, like data wrangling (data as unruly beasts to be tamed?) and munging (what is that, anyway?), neither of which mean much to me. I also like *data curation* but that's also a bit vague. Data carpentry probably has something to do with wishing I could make things like Carrie Roy, but I should start by saying what I don't like about the "data cleaning" or "janitor work" terms. To me these imply that there is some kind of pure or clean data buried in a thin layer of non-clean data, and that one need only hose the dataset off to reveal the hard porcelain underneath the muck. In reality, the process is more like deciding how to cut into a piece of material, or how much to plane down a surface. It's not that there's any real distinction between good and bad, it's more that some parts are softer or knottier than others. Judgement is critical.

The scale of data work is more like woodworking, as well. Sometimes you may have a whole tree in front of you, and only need a single board. There's nothing wrong with the rest of it, you just don't need it right now.

Finally, as the article points out, data work is as much about joining things together as it is selecting and pruning. Like building a data chair — you turn a dataset on the data lathe, and then glue it to the appropriate slot in another dataset. Carpentry has all these aspects, from selecting and shaping to careful joinery.

But there are other aspects of carpentry that make it an appealing metaphor. I don't know as much as I'd like about woodworking, but my impression is that it is not so much a single discipline as a vast array of specific skills. None of these are particularly difficult by themselves, but knowing which tool or method to use at each stage and carrying out each one cleanly and efficiently takes years of practice. Data carpentry, which I've been practicing in one way or another for about 15 years (though never as my official responsibility), is likewise not a single process but a thousand little skills and techniques. Instead of feeling the grain of a pine board, I spot the distinctive marks of broken UTF-8. Instead of a router and drill press I use perl (that's right, perl) and shell scripts. But even after that much time, I would still say I'm just fairly good. Building up solid instincts and a suite of processes makes me a lot faster than beginners, but I can't foresee running out of scripts to write.

So where does this leave us? I got the impression that the main purpose of the NYT article was to introduce a class of promising startups. I think these will do a good job of knocking off 60-80% of common cases, which will certainly save a lot of people a lot of time. I also think that if anyone can systematize data work, Jeff Heer and his team are a good bet (although writing an article about data work and not mentioning Hadley Wickham at RStudio is like writing an article about symphonies and not mentioning Beethoven). But I cannot imagine that data carpentry will not be a major part of the work of data science in the future. Every data set has its idiosyncrasies. You can streamline the process, but you can't avoid it. To draw out the analogy a bit more: sure, there's Ikea, but the best furniture is still made by Amish carpenters.

Most importantly, I don't think we want to make data carpentry automated — and therefore invisible. Lillian Lee and I were recently commiserating about the tendency of machine learning students to never look at the data they're working with. We both felt that to be really effective, you have to understand both the generalities and the specifics of a model. There's no substitute for experiencing the data set directly, and comparing this experience to a skilled, hands-on craft like woodwork feels right to me.

I teach Computer and Information Science at Cornell in Ithaca, NY. I work on text mining and machine learning, particularly unsupervised topic modeling and latent Dirichlet allocation. I like applications in computational approaches to history, literature, and social science. I tweet @dmimno.