Wheels of All Colors, Shapes, and Sizes
Working Together: A Vision of Common Purpose in Data Science

At Patterns, it doesn’t really matter what domain the data being discussed and used in our articles are from, what matters is that the data science solutions being proposed in the articles can be shared across domain boundaries. Often, when I am talking to people about Patterns, I describe the situation like this: everyone is creating data all the time, and we’re all coming across the same sorts of problems, like how to deal with massive data volumes, complicated analysis procedures, how to visualize the data, etc.

To look at it in another way: we’re all inventing wheels, but all the wheels we’re inventing are different colors, slightly different shapes, and with different axle fittings, so it’s really difficult to take these many wheels and make them work together as one part of a larger machine.

Patterns’s aim is to share knowledge of the wheels (i.e., data science solutions) that already exist, so researchers don’t have to design and develop them from scratch. Pushing the analogy even further, we also want to share the different ways wheels can be used. We’re not just talking car, truck, or bicycle wheels here—we’re also talking about flywheels, cogwheels, measuring wheels, Ferris wheels, mill wheels, even LPs and spinning disk hard drives, with all the many and varied uses they can be put to.

Just like wheels, some data science solutions come with their own built-in infrastructure (e.g., a Ferris wheel) that enables them to be used effectively as soon as they’re installed. Some other solutions are more like car or bicycle wheels, in that they require supporting infrastructure and hardware for them to operate. Just having a bicycle wheel is no use, if you have no frame to put it in or roads to cycle it on. And for some users, data science techniques are at that stage of development where they’re a bit like a unicycle, in that it’s possible for skilled operators to use them with a great deal of practice, but most people would find them very difficult to manage. For all of these techniques, we want to encourage their development to make them more user friendly and accessible, across the spectrum of their original research domain and beyond.

Wheels are a simplistic (if amusing) analogy. If anything, data and data science solutions can come in a far wider and more bewildering array of shapes and formats than wheels do. People can generally agree that a wheel is circular and turns around an axle, even if the details of what the wheel actually does vary. Data science techniques are far more nebulous and difficult to define.

If you want to stop a data scientists’ meeting in its tracks, one way to do so is to ask what data actually are. You won’t get a definitive answer, and might get a complaining response, because this question has been debated at great length by researchers over the years. For most people, data, like a wheel, are something that they can recognize when they see them, although when one person thinks about data, they may well be picturing something different from what another does.

This fundamental communication gap is one that needs to be closed when communicating across domain boundaries, just so that we’re not talking at cross-purposes. Data to an astronomer may well be at a completely different scale and level of complexity to that of a zoologist. Even within the data science community, what constitutes Big Data is hotly contested.

So, what do the dictionaries say data are? The Cambridge Dictionary defines data as “information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer” (https://dictionary.cambridge.org/dictionary/english/data), while Merriam-Webster defines data as “1. factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation; 2. information in digital form that can be transmitted or processed; 3. information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful” (https://www.merriam-webster.com/dictionary/data).

Whereas DataCite defines a dataset as “Recorded information, regardless of the form or medium on which it may be recorded including writings, films, sound recordings, pictorial representations, procedural manuals, forms, diagrams, work flow, charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data” (https://datacite.org/documents/Business_Models_Principles_v1.0.pdf).

There are commonalities between these (and other) definitions, in particular the idea of data being collected, examined, and used to inform decision-making. This is unsurprising, because this is fundamentally what we use data for, to help us understand a situation and make decisions based on that understanding. Another common theme in the definitions is the use of computers and digital forms, although not all data were born digital and not all research results are capable of being stored in a computer. (Examples of a physical data source would be a fossil, or an ice core that has captured the atmospheric gas composition of our atmosphere over many millennia.)

Providing a definitive answer to the question “what are data” is not something that can be done in the space of an editorial. Perhaps a better question to ask is “what are data to you?” To a researcher in computer vision, data are large collections of photographs harvested from the Internet with associated tags labeling them. To a meteorologist, data are the streams of numbers coming from their instruments, measuring things like air pressure, temperature, relative humidity, at regular intervals over the course of the day. To a business analyst, data are spreadsheet of IP addresses, along with the number of click-throughs produced as the result of an email marketing campaign. All these things are used by their collectors to inform
decision-making, and all these need to be carefully processed and analyzed in order to make useful decisions.

It has been said many times before that good decisions need good data. In the case of science, the data are the foundation of the conclusions, and if the conclusions are to hold, the data must be solid.

When it comes to publishing scientific results, the question of “what data did I create” becomes an important one, along with the definition of what data are for that particular domain. The definition is especially important for Patterns, as we are communicating across domain boundaries, and what goes without saying in one domain needs to be explicitly spelled out for users from other domains.

It’s for those reasons of easing communication, transparency, and verifiability that Patterns mandates a data and code accessibility statement, allowing readers to find the data and code that underpin the published research and learn more about them. It’s for these same reasons that we encourage authors and readers to think a little deeper about what data are for them and to start considering data and code to be important, first-class research outputs, worthy of sharing and discussion.

Data are not easy to define, and data science solutions to problems are not easy to standardize, but the benefits of common definitions and standardization are well understood. With Patterns, we’re aiming to improve cross-domain communication about data and data science in such a way as to be able to build better wheels, and solve more problems in an effective and understandable way.

WEB RESOURCES
“data,” Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/data
“data,” Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/data
Business Models Principles, DataCite, https://datacite.org/documents/Business_Models_Principles_v1.0.pdf

Sarah Callaghan
Editor-in-Chief, Patterns
https://doi.org/10.1016/j.patter.2020.100088