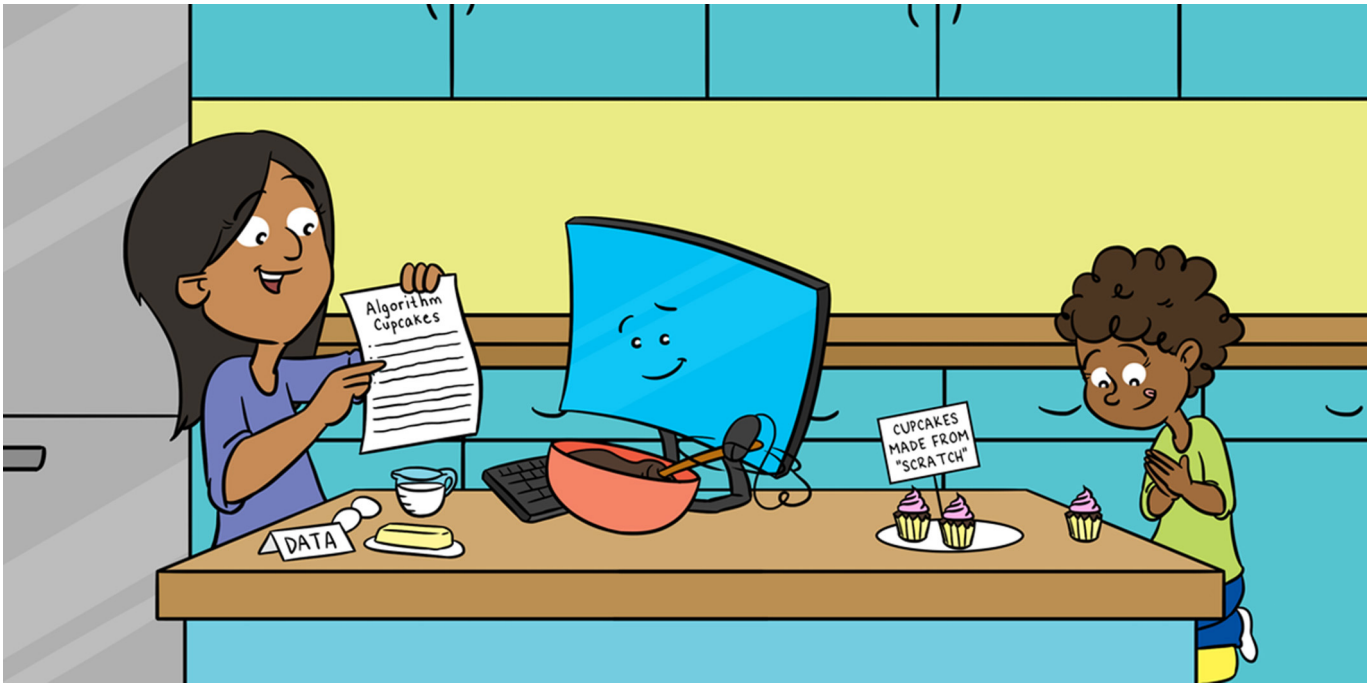
frontiers FOR YOUNG MINDS

MATHEMATICS
Published: 15 April 2021
doi: 10.3389/frym.2021.632923



COOKING UP KNOWLEDGE FROM BIG DATA USING DATA SCIENCE

Andrea Rau *

Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France

YOUNG REVIEWER:

JASMINE

AGE: 11

Data collected in very large quantities are called big data, and big data has changed the way we think about and answer questions in many different fields, like weather forecasting and biology. With all this information available, we need computers to help us store, process, analyze, and understand it. Data science combines tools from fields like statistics, mathematics, and computer science to find interesting patterns in big data. Data scientists write step-by-step instructions called algorithms to teach computers how to learn from data. To help computers understand these instructions, algorithms must be translated from the original question asked by a data scientist into a programming language—and the results must be translated back, so that humans can understand them. That means that data scientists are data detectives, programmers, and translators all in one!

DATA, DATA, EVERYWHERE

Data are a collection of information—numbers, measurements, words, or descriptions—that have been gathered and stored for a specific reason. Recently, many new tools have been developed that have made it quite easy to collect extremely large amounts of data. When data are available in huge quantities, they are often called **big data**. Big data have changed the way we think about and answer many very different questions, for example predicting the weather, finding routes to avoid getting stuck in a traffic jam, or suggesting a new television series you might like based on previous shows you watched.

BIG DATA

Extremely large and complex datasets that are challenging to store, process, analyze, and interpret. Data scientists often need to use specialized tools and methods to work with big data.

BIG DATA: A BIG CHALLENGE IN BIOLOGY!

Big data have also helped advance research in biology, which is the study of living things like humans, animals, plants, and bacteria. Many very specialized tools now allow big biological data to be collected in research laboratories, hospitals, out in nature, and even at home! For example, wearable devices like smart watches can have real-time sensors to help a doctor monitor how well you sleep. Drones can fly over farms and take pictures of fields to give a bird's-eye view of how the crops are doing. New laboratory techniques can now easily read a person's complete set of genetic instructions, made up of about 3 billion letters (to give you an idea of the scale, 3 billion seconds equals about 90 years!). With all this information available, it is a challenge to store, process, analyze, and interpret data, and we need computers to help.

DATA SCIENCE

Interdisciplinary field that combines tools from statistics, mathematics, and computer science to find interesting patterns from complex datasets, including big data.

MATHEMATICS + STATISTICS + COMPUTER SCIENCE + BIG DATA = DATA SCIENCE

Big data are so big that they have led to the development of a relatively new and exciting field called **data science**. Data science combines tools from many other fields, including statistics, mathematics, and computer science, to find interesting patterns from complex data. Data scientists must spend a lot of time organizing data before they can get to work. To answer a specific question, a data scientist needs to find or create a **dataset**, or a collection of datasets. Some datasets are publicly available for anyone to use, and a search engine like Google Dataset Search[1] can help you find one using keywords. Other datasets, like those including personal medical information, may only be available to a limited set of people. A data scientist might even need to collect new data to answer a question. For example, if you want to know the favorite colors of your classmates, you could write a survey to collect answers from the other students.

DATASET

A structured collection of related information—numbers, measurements, words, or descriptions—that has been gathered and stored for a specific reason.

[1] https://datasetsearch.research.google.com

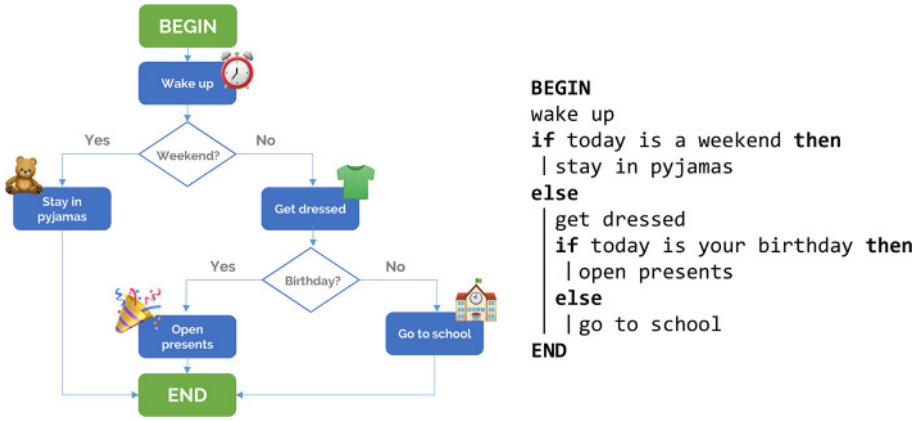


Figure 1

Figure 1

An algorithm is a set of step-by-step instructions for a computer. A useful way to visualize and build an algorithm is to draw a flowchart to connect each step to another. In flowcharts, rectangles might represent actions, and diamonds a decision. In the morning, you could use a flowchart like the one on the left to decide whether you can stay in pajamas, open birthday presents, or go to school. After drawing a flowchart, you could then translate the steps for your algorithm into a more detailed description, as shown on the right.

FROM MESSY TO TIDY DATA

A big part of a data scientist's job is to get the data they want to use into a useable format. One way to think about this is to imagine big data as a jumble of all your LEGOs® scattered all over your home. Before you can start sorting through your blocks to build something, you must do some tidying and get them all into a pile in the same room! Most real datasets are very "messy," meaning that they might include typos or even missing values. As an example, some responses to your survey on favorite colors might include "blue," "Blue," "BLUE," and "Bluue." To make these data easier to understand, you would need to tidy the data by changing all these variations to a single value, like "blue," as they all mean the same color.

ALGORITHMS AS RECIPES FOR DATA SCIENCE

Once your LEGOs® are all in one spot, there are lots of goals you might have, for example grouping blocks into sets, or predicting the kind of set you might like next. If you have a small number of LEGOs®, it might be easy to do this by hand—but for big data, we need special tools to help. One powerful tool for dealing with big data is called **machine learning**, which is when we teach a computer how to learn from data without first giving it the answer. To do this, data scientists must give the computer a set of detailed, step-by-step instructions, called an **algorithm** (Figure 1). These instructions must be written in a way that the computer can understand, and this is called **coding**. You can think of an algorithm like a recipe for baking a cake. The recipe starts with a set of ingredients (your data), and it tells you exactly how (your algorithm) to mix the batter, heat the oven, and bake it to get a tasty dessert (your results). The difference between a recipe and an algorithm, though, is that the instructions of an algorithm must be very precise so that the computer knows exactly what to do. In a recipe, instead of saying "mix in a dash of salt to the batter," it would be like

MACHINE LEARNING

The use of algorithms to teach a computer how to automatically learn from data and improve from experience without help from a human.

ALGORITHM

A set of detailed, step-by-step instructions or rules to be followed by a computer.

CODING

Using a programming language to communicate with a computer and provide it with instructions, referred to as an algorithm.

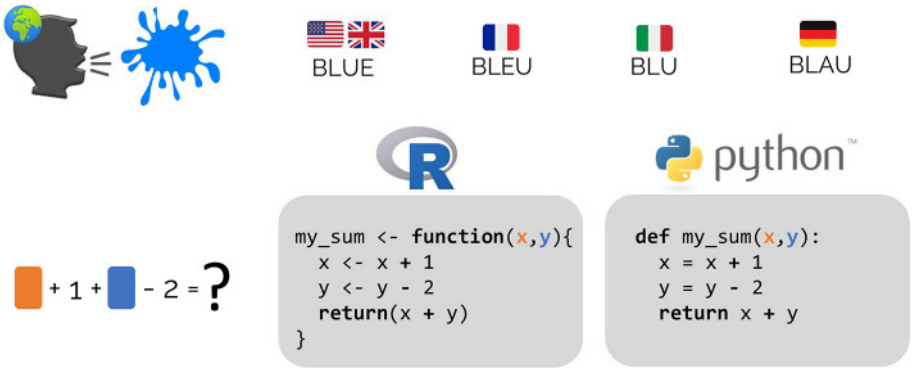


Figure 2

Figure 2

Algorithms can be coded using different coding languages, just as ideas can be expressed using different languages. Let us say we want to write an algorithm that would take any two numbers, add 1 to the first and subtract 2 from the second, and then add them together. If we start with 2 and 4, we want to teach the computer to give us $(2 + 1) + (4 - 2) = 5$ as an answer. Our algorithm, which we called my_sum, looks similar in the R and Python coding languages, but if you look closely, you can see some differences.

saying, "add 1 gram of salt to the batter and stir three times with a wooden spoon."

WHAT LANGUAGE DO BOTH YOU AND YOUR COMPUTER SPEAK?

Coding is a way to translate a scientific question into a language your computer can speak. There are many different languages spoken by people all over the world (English, French, Italian, German, ...), and similarly, there are many different coding languages that can be used to write an algorithm (Figure 2). Just as a recipe written in English and French might say the same thing in two unique ways, different coding languages put together instructions for the computer in different ways. New coding languages are invented every year! There is even a coding language that was created especially for kids who are 8–16 years old, called Scratch[2] [1]. Two popular coding languages that are often used by data scientists today to write algorithms are called R and Python. Both languages are **open source**, which means that data scientists who write their algorithms in these languages can share them with everyone for free. This makes it easy for data scientists to work together and help improve each other's code!

[2] https://scratch.mit.edu

OPEN SOURCE

Type of computer software that is community developed and supported. Open source code and software are typically free for anyone to use, share, and modify.

COMBINING COMPUTER RECIPES INTO A DATA SCIENCE COOKBOOK

A data scientist might have to write several algorithms and combine them to get the answer they are looking for. Just as a chef might collect several recipes together in a cookbook, a data scientist sometimes creates or uses bundles of algorithms, called **software packages**. When software packages are written in an open-source language like R or Python, that can help data scientists create reproducible work. Reproducible data science means that other people can easily re-run, repeat, and reuse a scientist's work. This helps everyone work more efficiently and easily share what they find with others. Reproducibility

SOFTWARE PACKAGE

An organized collection of related algorithms that work together for a particular task or have a similar function.

also helps to build trust that the algorithms are correct. In the same way, you can give your favorite cookbook to a friend so they can make that tasty cake for themselves!

CONCLUSIONS

Big data keep getting bigger, whether in biology, banking, or marketing, and big data will continue to have a huge impact on our lives. However, there are also growing concerns about the consequences of big data collection on privacy. When you sign up for a free service or app (like social media, email, video streaming, or location-sharing services), in exchange you agree to let a privately owned company collect data about you. That data might include the keywords you search for, the websites you browse, the videos you like, or the places in your neighborhood that you visit. Companies use that data to create advertising targeted specifically to you, often with the goal of selling you as much as possible! You can take steps to be aware of what types of data are being collected about you, for example by looking within the settings of the apps. This can help you limit the collection of some types of data, like your location information, and also help you to decide which apps and services you trust and which you should consider uninstalling.

In the coming years, we will need lots of new data scientists who can help make sense of big data with machine-learning methods. It will be especially important for people from all different backgrounds to help make sure that everyone can benefit equally from these analyses. It is an exciting time to be a data scientist—we are like detectives, mathematicians, artists, computer programmers, and translators, all rolled into one!

REFERENCES

1. Maloney, J., Resnick, M., Rusk, N., Silverman, B., and Eastmond, E. 2010. The scratch programming language and environment. *ACM Trans. Comput. Educ.* 10:1–15. doi: 10.1145/1868358.1868363

SUBMITTED: 24 November 2020; **ACCEPTED:** 19 March 2021;
PUBLISHED ONLINE: 15 April 2021.

EDITED BY: Norma Ortiz-Robinson, Grand Valley State University, United States

CITATION: Rau A (2021) Cooking Up Knowledge From Big Data Using Data Science. Front. Young Minds 9:632923. doi: 10.3389/frym.2021.632923

CONFLICT OF INTEREST: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed

as a potential conflict of interest.

YOUNG REVIEWER

JASMINE, AGE: 11

My name is Jasmine. I like strategy-based cooperative board games. I also like reading Agatha Christie and other classic detective novels. I enjoy skiing, swimming, kayaking, and martial arts. I am a black belt in karate, which is my favorite accomplishment because it was very challenging and it took 6 years to finish the curriculum.

AUTHOR

ANDREA RAU

I am a biostatistician, data scientist, and researcher at the French National Research Institute for Agriculture, Food and Environment (INRAE) in Jouy en Josas, France. I develop statistical models and write computer code to help biologists find interesting patterns in their genomics data. In addition to writing computer code in a programming language called R, I speak both English and French at work. In my free time, I love cooking up new recipes and playing with my daughter Elise and my dog Bella. *andrea.rau@inrae.fr