IN-DEPTH

Building the Best Data Science Toolkit: Programming Languages

By David Ramel 02/22/2018

Pick the right tool for the job, they say. So what are the right tools for the coveted, mysterious and highly paid position of data scientist?

To find out, we checked in with our resident data scientist/tutorial author, interviewed the head of a sports analytics data science team, scoured the Web's many takes on the subject and gauged

the sentiment in various coding forums where that question has been asked repeatedly over the years.

We'll start off with this examination of programming languages and subsequently will investigate data science libraries and finally look at other miscellaneous tools.

The subject of the optimal data scientist toolkit came up this week -- again -- in a Hacker News **post**, where the conversation -- as it often does -- turned to programming languages.

The choice of a programming language often influences what other tools and libraries are used in a particular setup, and in the data science world, it basically comes down to using **R** or **Python**. Or a combination of the two. Or some other language. Or some other combination of languages.

Yes, in the nebulous world of data science, reaching a consensus on what programming language to use is about as hard as defining what a data scientist even is.

"What does 'Data Scientist' actually mean these days?" says a comment on the Hacker News thread. "Does it mean 'Write 10 lines of Python or R, and not fully understand what it actually does'? Or something else?"

A reader who hires data scientists provided the answer: "It means someone who can work with business stakeholders to break down a problem e.g. 'we don't know why customers are churning', produce a machine learning model or some adhoc analysis (usually the former) and either communicate the results back or assist in deploying the model into production. Typically there will be data engineers who will be doing acquisition and cleaning and so the data scientists are all about (a) understanding the data and (b) liaising with stakeholders."

With that settled, the reader noted the technologies typically used: "R/Python with Spark/H20 on top of a data lake i.e. HDFS, S3."

That data science expert, like many, sees a role for both R and Python. For example, another data scientist on the thread said he uses Python for general purpose programming, R for statistics, bash for cleaning up files and SQL for querying databases.

"When I put on my data science/machine learning hat, I use Python, which has clearly established itself as the dominant language for DS/ML."

Dr. James McCaffrey, Microsoft Research

For our in-house expert, Dr. James McCaffrey, the choice is clearly Python, as any regular reader of sister publication *Visual Studio Magazine* knows from his series of **Python-based tutorials** dealing with everything from neural network time series regression to neural network L2 regularization.

"When I write C# code, I use Visual Studio," said McCaffrey, who works for Microsoft Research. "But when I put on my data science/machine learning hat, I use Python, which has clearly established itself as the dominant language for DS/ML."

Earlier, we interviewed Dr. Patrick Lucey, director of data science at STATS LLC, for his take on the matter. He heads a team of Ph.Ds at the prominent sports analytics company that counts some of the most popular sports teams and biggest media companies in the world among its customers. "At STATS, we use a varied number of languages -- our predominant one for prototyping is Python with Sci-kit learn (Numpy, Scipy and Matplotlib)," Lucey **said**.

"At STATS, we use a varied number of languages -- our predominant one for prototyping is Python with Sci-kit learn (Numpy, Scipy and Matplotlib)."

Dr. Patrick Lucey, STATS LLC

So both McCaffrey and Lucey are in the Python camp, but that's by no means universal. The Coursera educational site, for example, offers a **course** called "The Data Scientist's Toolbox," wherein R is taught, along with associated tools like **RStudio**.

While McCaffrey and Lucey both favor Python, an HN reader made the argument that both R and Python -- along with **Julia** -- are all Turing-complete languages, so two could be dropped from the mix and a data scientist could get by with just one of them.

Julia also got some love from other readers. "As a data scientist who has been using the language for 5 years now, Julia is by far the best programming language for analyzing and processing data," one said. Julia, probably not as well-known as R or Python, is described as "a high-level, high-performance dynamic programming language for numerical computing."

It's nearly certain, however, that the data scientist's toolkit for 2018 will include R or Python -- and probably both. R and Python, for example, are listed No. 1 and No. 2 in an article titled "**The Data Science Toolkit: 24 Free Data Science Tools**" on the Springboard educational site.

Here are the Springboard descriptions of each language:

- *R is a programming language used for data manipulation and graphics. Originating in 1995, this is a popular tool used among data scientists and analysts. It is the open source version of the S language widely used for research in statistics. According to data scientists, R is one of the easier languages to learn as there are numerous packages and guides available for users.*

- *Python is another widely used language among data scientists, created by Dutch programmer Guido Van Rossem. It's a general-purpose programming language,*

> *focusing on readability and simplicity. If you are not a programmer but are looking to learn, this is a great language to start with. It's easier than other general-purpose languages and there are a number of tutorials available for non-programmers to learn. You can do all sorts of tasks such as sentiment analysis or time series analysis with Python, a very versatile general-purpose programming language. You can canvass open data sets and do things like sentiment analysis of Twitter accounts.*

On the Reddit social coding site, an Ask-Me-Anything (AMA) **event** with data scientist Jake Porway of DataKind revealed he used both languages.

"I cut my teeth in the statistics department at UCLA, so when it comes to the analytical side of things I'm R all the way," Porway said. "For most other things (scripting, munging, etc.) Python is my tool of choice. Given its relatively shallow learning curve and increasingly robust data stack, I'd say Python is a great tool to start with and pretty much the lingua franca of data scientists these days. I can't really think of many others that are as 'simply useful, no matter what the task.'"

So, to get in on the action and become a successful data scientist -- last year named the **"best job in America**" by careers site **Glassdoor** and earlier named "**the sexiest job of the 21st century**" by *Harvard Business Review* -- definitely brush up on your R and Python skills.

Next time, we'll look at the best libraries for your modern data scientist toolkit. Stay tuned.

About the Author

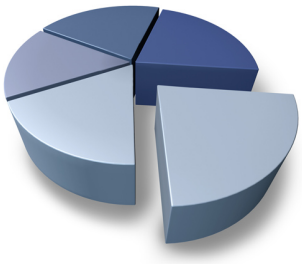
David Ramel is the editor of Visual Studio Magazine.

PRINTABLE FORMAT

Featured

Jakarta: The Community Reacts to a New Brand with an Old Pedigree

Stack Overflow Developer Survey: Dangers of AI, DevOps, Ethics, Python, More



Oracle Plans to Decouple JavaFX from the JDK

Most Popular Articles

Free White Papers

Data Governance Implementation Survey 2018

MORE TECH LIBRARY

Upcoming Events

Visual Studio Live! Austin
April 30-May 4, 2018
Austin, TX

Visual Studio Live! Boston
June 10-14, 2018
Cambridge, MA

Visual Studio Live! Redmond
August 13-17, 2018
Redmond, WA

Visual Studio Live! Chicago
September 17-20, 2018
Chicago, IL

Visual Studio Live! San Diego
October 7-11, 2018
San Diego, CA

Live! 360 Orlando
December 2-7, 2018
Orlando, FL