Hadley Wickham, the Man Who Revolutionized R

Jul 24, 2015 · 12,705 views

Share

T

"Fundamentally learning about the world through data is really, really cool."

~ Hadley Wickham, prolific R developer

***

If you don't spend much of your time coding in the open-source statistical programming language R, his name is likely not familiar to you -- but the statistician Hadley Wickham is, in his own words, "nerd famous." The kind of famous where people at statistics conferences line up for selfies, ask him for autographs, and are generally in awe of him. "It's utterly utterly bizarre," he admits. "To be famous for

writing R programs? It's just crazy."

Wickham earned his renown as the preeminent developer of packages for R, a programming language developed for data analysis. Packages are programming tools that simplify the code necessary to complete common tasks such as aggregating and plotting data. He has helped millions of people become more efficient at their jobs -- something for which they are often grateful, and sometimes rapturous. The packages he has developed are used by tech behemoths like Google, Facebook and Twitter, journalism heavyweights like the *New York Times* and FiveThirtyEight, and government agencies like the Food and Drug Administration (FDA) and Drug Enforcement Administration (DEA).

Truly, he is a giant among data nerds.

***

Born in Hamilton, New Zealand, statistics is the Wickham family business: His father, Brian Wickham, did his PhD in the statistics heavy discipline of Animal Breeding at Cornell University and his sister has a PhD in Statistics from UC Berkeley.

If there is such a thing as a data structure prodigy, Wickham might be one. His experience, he tells us with pride, started early on:

> *"My first job, when I was 15, was developing Microsoft Access Databases. I found that kind of fun. I was doing database documentation. They are still using the database I wrote."*

During this first job, Wickham began to reflect on better ways to store and manipulate data. "I've always been very certain that I could come up with a good way of doing things," he explained, "and that that way would actually help people." Although he didn't know it at the time, he believes it was then that he "internalized" the concept of Third Normal Form, a database design concept that would become central to his future work. Third Normal Form is essentially a manner of structuring data in a way that reduces duplication of data and ensures consistency. Wickham refers to such data as "tidy," and his tools promote and rely on it.

The logo for R. The programming language Hadley Wickham helped revolutionize.

Wickham first encountered the programming language R as a statistics major at the University of Auckland in New Zealand. Wickham describes R as "a programming language for understanding data."Along with SQL and Python, it is one of the most popular programming languages for data scientists.

Like Wickham, the programming language he would come to transform is from New Zealand. R was created in 1993 at the University of Auckland by statisticians Ross Ihaka and Robert Gentleman. The language was designed for data analysis, and has some quirks (like the way data structures are indexed and have to be stored in physical memory), so programmers coming from other languages often find it peculiar. Having programmed in Java, VBA, and PHP, Wickham found R to be "totally different." "[Many programmers] see R and think it is ridiculous and awful, but that didn't happen to me," he says. "I thought it was really interesting."

It wasn't until he he moved to the U.S., to start his PhD program at Iowa State University, that Wickham began developing R packages. In Wickham's words, making a package entails writing "some code that helps people solve problems -- and then you have to document it, so others can understand how to use it." The first package he created, as part of a class project, was for visualizing bioinformatic data. Though the package was never released to the public, he loved the idea of sharing his methods.

In 2005, he released the reshape package, the first of his many packages that would become a hit. The package has since been downloaded hundreds of thousands of times. The goal of reshape was to make aggregating and manipulating data less "tedious and frustrating." Easing the process of transforming data may not seem like a big deal to non-programmers, but for many data scientists and statisticians, this can often be the most time-consuming part of their work.

Wickham clearly relished the success of reshape. He developed the package because he didn't think the options available were good enough. Though not a braggart, Wickham does not lack in the confidence department. "I have a very strong belief that I know the right way to do things," he reiterates, "for better, or for worse."

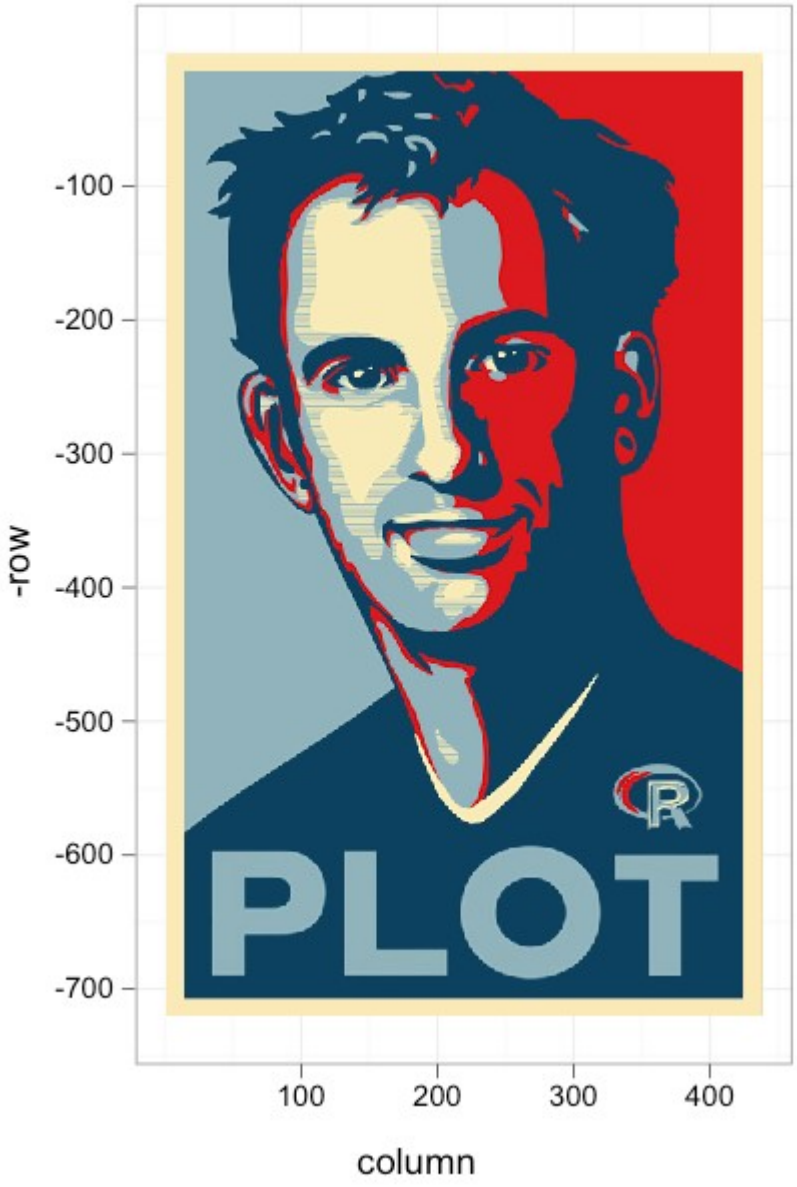
***

At the same time that reshape, and several other of his packages, were being widely used, Wickham was becoming disenchanted with the statistics discipline. He observed during his statistics PhD that there was a "total disconnect between what people need to actually understand data and what was being taught." Unlike the statisticians who were focused on abstruse ramifications of the central limit theorem, Wickham was in the business of making data analysis easier for the public. He expands on this:

> *"There are definitely some academic statisticians who just don't understand why what I do is statistics, but basically I think they are all wrong . What I do is fundamentally statistics. The fact that data science exists as a field is a colossal failure of statistics. To me, that is what statistics is all about. It is gaining insight from data using modelling and visualization. Data munging and manipulation is hard and statistics has just said that's not our domain."*

In the middle of that period of frustration, Wickham developed **ggplot2**. Downloaded millions of times, it would become, by far, his most popular package, and would change the way many people

conceptualized data visualization. ggplot2's success would also allow Wickham to leave academia, and focus exclusively on improving R, through his post as the Chief Scientist at RStudio (the for-profit creators of the most popular integrated development environment (IDE) for R).



Hadley Wickham placed on a graph created with ggplot2; Image by David Kahle and Garrett Grolemund

ggplot2 was developed based on statistician Leland Wilkinson's "Grammar of Graphics", a formalization of data visualization theory. Wickham sees ggplot2 and the grammar of graphics as "a way of thinking about visualization not as a series of mechanical operations (draw a line from here to there, draw points here, color a rectangle here) but instead thinking about visualization as a way mapping data to things that you can see."

The concepts behind the grammar of graphics are relatively abstract. The big idea is that charts are made of “geometries” (the graphical element you see on the chart like a point or a bar), and “aesthetics” (the directions concerning where the geometries are to be placed). It may not sound revolutionary, but the implementation of this concept created by Wickham has made the process of making charts easier for hundreds of thousands of people. The question-and-answer website Stack Overflow, which has nearly 9,000 questions tagged as questions concerning ggplot2, even refers to ggplot2 as making graphics in R “fun”. Charts made with ggplot2 have appeared in Nature, FiveThirtyEight, and The New York Times.

priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/

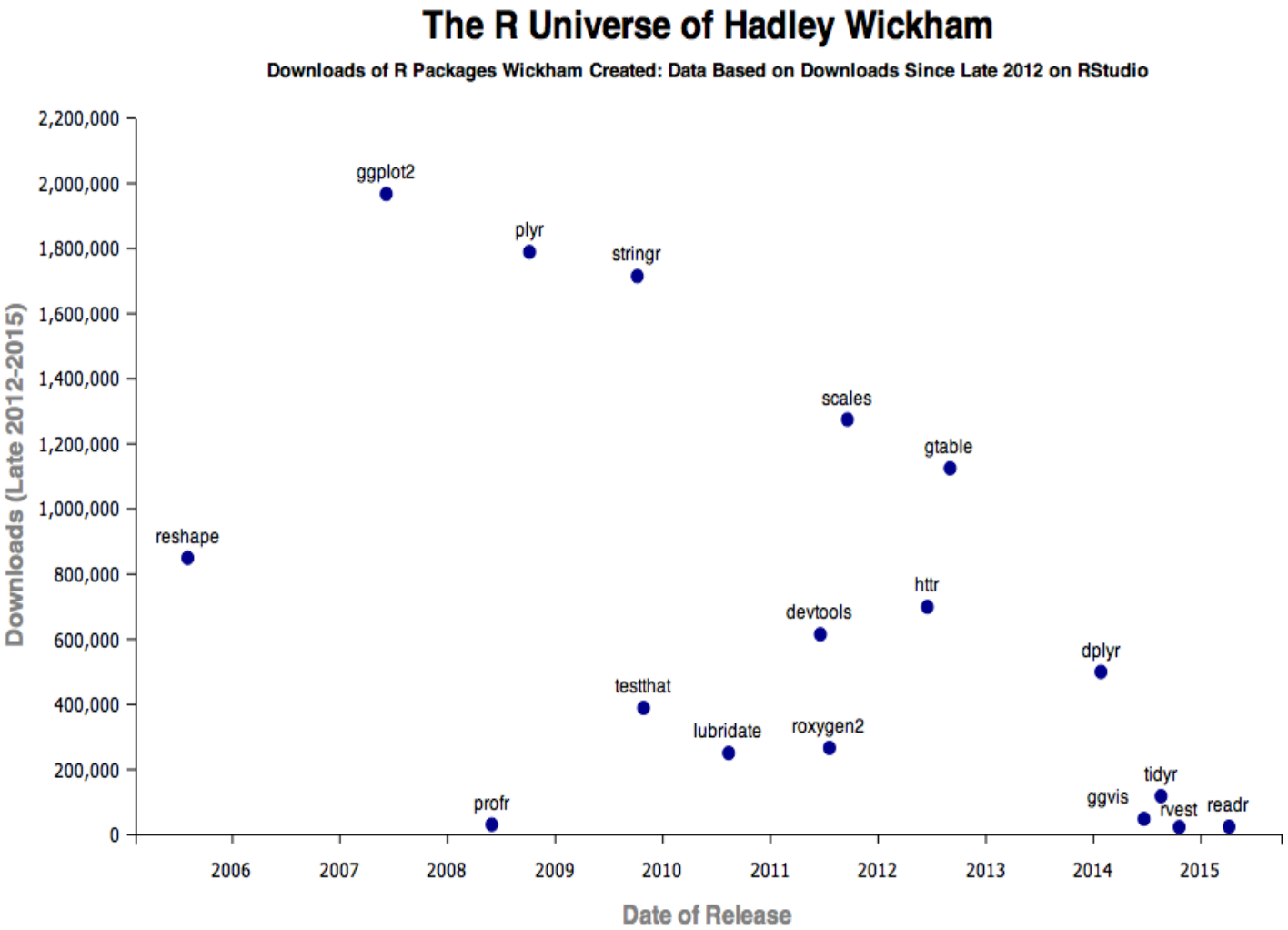
Hadley Wickham holding a Chinese Translation of a book on his visualization package ggplot2; Image from statr

In addition to developing ggplot2 and reshape, Wickham has designed a number of other wildly popular packages to solve other major problems for data scientists. Need to easily manipulate data in

the form of words (strings)? Want to scrape data from the web? Need to easily write your own package? Wickham has you covered.

On Quora, one R users asks: "How is Hadley Wickham able to contribute so much to R, particularly in the form of packages? I still can't wrap my head around how much Hadley produces. It doesn't seem possible..." Eduardo Arino de la Rubia, an active member of the R community, suggests that all successful programming languages need "luminaries" like Hadley. He compares him to David Heinemeier Hansson (the creator of the web application framework Ruby on Rails) and Tatsuhiko Miyagawa (an important developer for the programming language Perl).

The chart below displays the initial release date and number of downloads for 17 of Hadley's packages that have been downloaded over 20,000 times (sometimes referred to as the Hadleyverse). The downloads numbers are bare minimums, because they only reflect downloads from one popular download source since late 2012. And yes, this chart was made using one of Hadley's packages (ggvis):



Dan Kopf, Priceonomics; Data: cranlogs

So why create all of this? R is free to download, as are all of his packages, so the financial incentive is minor. Simply put, it rankles him when a problem is more difficult to solve than it should be. While "most other people accept that life is hard", Wickham does not.

"One of the attributes that has made me successful," he says, "is that I am exquisitely sensitive to frustration."

This sensitivity has earned him a peculiar, underground level of fame.

In the vast majority of circumstances, Wickham is inconspicuous, but when he is at an R meetup or a statistics conference, he becomes a rock star. "I can see my level of fame getting to the point where it's actively uncomfortable," he says. He wishes someone would write a book about "how to be a celebrity in a very specific field," and worries that he doesn't know how to appropriately act when people are gushing over him.

Although now accustomed to the notoriety, he can still be thrilled by the uses of tools he created. He found it particularly cool to see how many people at "Facebook, Google, Twitter, Tumblr…" use his tools. He noted that only in San Francisco is there is a substantial probability of someone recognizing him on the street. He also recalled being delighted by a recent tour he took of data journalism media outlet FiveThirtyEight. He thought it was cool how much they used his tools (their graphics are created using a heavily customized version of ggplot2).

***

Above all else, Wickham is motivated by empowering people who like to play with data. "Fundamentally learning about the world through data is really really cool," he explains. "The analyses that get me excited are not Google crunching a terabyte of web ad data in order to optimize revenue… [but rather] the biologists who are absolutely passionate about this one swampfly and now they can use R and they can understand it."