I'm Hadley Wickham, Chief Scientist at RStudio and creator of lots of R packages (incl. ggplot2, dplyr, and devtools). I love R, data analysis/science, visualisation: ask me anything!

hadley [1] and r/Science AMAs[1]

[1]Affiliation not available

April 17, 2023

Abstract

Broadly, I'm interested in the process of data analysis/science and how to make it easier, faster, and more fun. That's what has lead to the development of my most popular packages like ggplot2, dplyr, tidyr, stringr. This year, I've been particularly interested in making it as easy as possible to get data into R. That's lead to my work on the DBI, haven, readr, readxl, and httr packages. Please feel free to ask me anything about the craft of data science. I'm also broadly interested in the craft of programming, and the design of programming languages. I'm interested in helping people see the beauty at the heart of R and learn to master it as easily as possible. As well as a number of packages like devtools, testthat, and roxygen2, I've written two books along those lines: Advanced R, which teaches R as a programming language, mostly divorced from its usual application as a data analysis tool. R packages, which teaches software development best practices for R: documentation, unit testing, etc. Please ask me anything about R programming! Other things you might want to ask me about: I work at RStudio. I'm the chair of the infrastructure steering committee of the R Consortium. I'm a member of the R Foundation. I'm a fellow in the American Statistical Association. I'm an Adjunct Professor of Statistics at Rice University: that means they don't pay me and I don't do any work for them, but I still get to use the library. I was a full time Assistant Professor for four years before joining RStudio. These days I do a lot of programming in C++ via Rcpp. Many questions about my background, and how I got into R, are answered in my interview at priceonomics. A lot of people ask me how I can get so much done: there are some good answers at quora. In either case, feel free to ask for more details! Outside of work, I enjoy baking, cocktails, and bbq: you can see my efforts at all three on my instagram. I'm unlikely to be able to answer any terribly specific questions (I'm an amateur at all three), but I can point you to my favourite recipes and things that have helped me learn. I'll be back at 3 PM ET to answer your questions. ASK ME ANYTHING! Update: proof that it's me Update: taking a break. Will check back in later and answer any remaining popular/interesting questions

I'm Hadley Wickham, Chief Scientist at RStudio and creator of lots of R packages (incl. ggplot2, dplyr, and devtools). I love R, data analysis/science, visualisation: ask me anything!

HADLEY R/SCIENCE

ABSTRACT

Broadly, I'm interested in the process of data analysis/science and how to make it easier, faster, and more fun. That's what has lead to the development of my most popular packages like ggplot2, dplyr, tidyr, stringr. This year, I've been particularly interested in making it as easy as possible to get data into R. That's lead to my work on the DBI, haven, readr, readxl, and httr packages. Please feel free to ask me anything about the craft of data science.

I'm also broadly interested in the craft of programming, and the design of programming languages. I'm interested in helping people see the beauty at the heart of R and learn to master it as easily as possible. As well as a number of packages like devtools, testthat, and roxygen2, I've written two books along those lines:

Advanced R, which teaches R as a programming language, mostly divorced from its usual application as a data analysis tool.

R packages, which teaches software development best practices for R: documentation, unit testing, etc.

Please ask me anything about R programming!

Other things you might want to ask me about:

I work at RStudio.

I'm the chair of the infrastructure steering committee of the R Consortium.

I'm a member of the R Foundation.

I'm a fellow in the American Statistical Association.

I'm an Adjunct Professor of Statistics at Rice University: that means they don't pay me and I don't do any work for them, but I still get to use the library. I was a full time Assistant Professor for four years before joining RStudio.

These days I do a lot of programming in C++ via Rcpp.

Many questions about my background, and how I got into R, are answered in my interview at priceonomics. A lot of people ask me how I can get so much done: there are some good answers at quora. In either case, feel free to ask for more details!

Outside of work, I enjoy baking, cocktails, and bbq: you can see my efforts at all three on my instagram. I'm unlikely to be able to answer any terribly specific questions (I'm an amateur at all three), but I can point you to my favourite recipes and things that have helped me learn.

I'll be back at 3 PM ET to answer your questions. ASK ME ANYTHING!

Update: proof that it's me

Update: taking a break. Will check back in later and answer any remaining popular/interesting questions

👁 **READ REVIEWS**

✎ **WRITE A REVIEW**

CORRESPONDENCE:

Thanks for changing the way I use and program in R Hadley.

You've worked a lot on data ingest and visualisation. What are your thoughts on the future of modelling in R? Is there room for a comprehensive grammar-like DSL like dplyr and ggplot, dedicated to fitting models?

yoplaitful

Yes, absolutely! But I'm not entirely sure what a grammar of modelling should look like. I suspect it will be focussed around model building, not so much the mechanics of model building. I've been starting to explore a little what it might look like with purrr and dplyr, e.g. https://github.com/hadley/purrr#examples. I'm not exactly sure what the verbs should be, but I think the fact that you can put linear models in a data frame column to be profoundly important.

How would you teach a brand new student R? i.e. what do you think is a good pathway for them to go from a complete beginner to proficient?

Also what's your favorite type of bbq? And any fav bbq restaurants?

sarahbotts

I'd absolutely recommend starting with visualisation. It's great because creating a visualisation is a big payoff, and that's needed to help students work through the pain of learning a new (programming) language. Then you need to learn about data manip, tidy data, modelling, communicating results, ... I'm working on a book (with Garrett Grolemund) that will hopefully pull all these pieces together: http://r4ds.had.co.nz

I'd also recommend looking at project mosaic - the academics involved are very thoughtful about what's the minimal useful subset of R/statistics/data science you need to be useful. And I'd recommend reading Badass: making users awesome and thinking about how you can make students awesome.

I have a few other notes about teaching (in the short course scenario) at https://gist.github.com/hadley/37c8078eb9d46b5dac7e

Are there efforts underway to make it easier to integrate Python code from within RStudio? For instance, I love that rmagic makes it easy to call R code from iPython notebooks, but I primarily work in R and would like to see the opposite. More generally, do you have thoughts on the debates among data scientists as to which language is better as a primary data language?

p.s. Thank you for making data analysis and visualization so *ing easy. Dplyr is a godsend, and RStudio has made it possible for me to push all of the analysts in the city where I work away from excel and towards R.

oreo_fanboy

I think Python support in RStudio (the IDE) is gradually improving over time, but it's obviously not a focus of RStudio (the company). But we are thinking about notebooks...

Generally, I think R and python are much more similar than they are different. I'm not really interested in the debates about which one you should learn. Obviously, I think learning R is the right choice, but you can be effective with either. My main advice is to focus on one and get good at it. That's a much more effective way of learning than dabbling in both. (Of course, once you get good in one, you can learn the other, but do it in serial, not parallel)

Do you still hate secondary axes, and why so?

In 2011, you professed your profound dislike for seconday y-axis.

I'm not using ggplot2 because this feature is absent. Can I try again and give you two examples where they are useful?

- **Temperature plot with fahrenheit on the left axis and celcius on the right (one single line, two axes)**

- **Price of oil in USD/bbl on the left and in EUR/bbl on the right (two lines). This one could be rebased to 100, but we would be losing the actual units.**

neuro99

Yes, I still stand by that position. I agree that they can be useful when the axes are simple linear transformations of each other, but I don't think they're useful enough for me to spend hours to implement them.

With the amount of attention that the "Big Data" craze is getting and some of the limitations RStudio has when handling large amounts of data, do you foresee better server integration for the common man (by common man, I mean poor graduate student)?

arifyali

I've included my general big data thoughts from a recent interview below. It's hard to give specific advice without knowing more about your data.

Big data is extremely overhyped and not terribly well defined. Many people think they have big data, when they actually don't. I think there are two particularly important transition points:

- From in-memory to disk. If your data fits in memory, it's small data. And these days you can get 1 TB of ram, so even small data is big! Moving from in-memory to on-disk is an important transition because access speeds are so different. You can do quite naive computations on in-memory data and it'll be fast enough. You need to plan (and index) much more with on-disk data

- From one computer to many computers. The next important threshold occurs when you data no longer fits on one disk on one computer. Moving to a distributed environment makes computation much more challenging because you don't have all the data needed for a computation in one place. Designing distributed algorithms is much harder, and you're fundamentally limited by the way the data is split up between computers.

I personally believe it's impossible for one system to span from in-memory to on-disk to distributed. R is a fantastic environment for the rapid exploration of in-memory data, but there's no elegant way to scale it to much larger datasets. Hadoop works well when you have thousands of computers, but is incredible slow on just one machine.

Fortunately, I don't think one system needs to solve all big data problems. To me there are three main classes of problem:

1. Big data problems that are actually small data problems, once you have the right subset/sample/summary. Inventing numbers on the spot, I'd say 90% of big data problems fall into this category. To solve this problem you need a distributed database (like hive, impala, teradata etc), and a tool like dplyr to let you rapidly iterate to the right small dataset (which still might be gigabytes in size).

2. Big data problems that are actually lots and lots of small data problems, e.g. you need to fit one model per individual for thousands of individuals. I'd say ~9% of big data problems fall into this category. This sort of problem is known as a trivially parallelisable problem and you need some way to distribute computation over multiple machines. The foreach is a nice solution to this problem because it abstracts away the backend, allowing you to focus on the computation, not the details of distributing it.

3. Finally, there are irretreviably big problems where you do need all the data, perhaps because you fitting a complex model. An example of this type of problem is recommender systems which really do benefit from lots of data because they need to recognise interactions that occur only rarely. These problems tend to be solved by dedicated systems specifically designed to solve a particular problem.

Cheers for all your hard work for the greater good of R.

What do you foresee as the 'next big thing' in R development? For example, ggplot2 converted me to using R as it made graph building super intuitive and looks great. Anything on the horizon that you think might have a similar impact?

BooRadleyBoo
I think a grammar of modelling, e.g.https://www.reddit.com/user/yoplaitful, is really important.

But I think people tend to over emphasise the importance of revolution over evolution. I think it's just as valuable to spend my time continuously polishing the rough edges of R, so that all the little things just get easier and easier. I want you to spend your precious cognitive resources on the particular challenges of your data analysis, not fighting R to get it to do what you want.

Since this is DatalsBeautiful, what would you consider to be the most beautiful data visualization you've seen done with ggplot2?

zonination
There are a lot, but I think James Cheshire has done a lot of beautiful work. London: The Information Captial contains many beautiful graphics. Many are done with ggplot2.

I'm also broadly interested in the craft of programming, and the design of programming languages. I'm interested in helping people see the beauty at the heart of R and learn to master it as easily as possible.

I use R (and your packages) often, and find it extremely powerful and useful as a scientist. However, I come from a more traditional CS background, and have programmed in many languages, and honestly consider R to be one of the ugliest languages I frequently use (not trying to be offensive, just truthful). I know this is subjective, but it is also somewhat common for people who learn a lot of other languages before R. Do you understand why people might think it is ugly, and what perspective can you give that will let them see the beauty that you see? Also, (again purely subjective) I find julia to be the most elegant and beautiful language I know, what are your thoughts on it? Have you ever considered designing your own language from scratch, and if not, what massively breaking change in R would you have introduced from the outset if you had a time machine?

DrGar
To give you some perspective, the languages I have programmed the most in are VBA, PHP, R, and C++. These are all languages widely considered to ugly/awful/the worst programming language you have ever seen. But to me, all of these languages are incredibly pragmatic: they designed to solve a specific problem, not to appeal to some abstract/pure vision of beauty.

The better I understand R, the more I appreciate the vision of John Chambers, Ross Ihaka, and Rob Gentleman. Many of the features of R that seem quirky at first, I think are actually well tailored to the problem of data analysis. (Of course, there's lots of mistakes and bad code in base R, but I think the language itself is quite elegant).

I think you also need to be quite careful with aesthetic judgements - it's hard to separate out what is truly ugly from what is just new (to you). When ever you feel visceral revulsion towards something, you need to check that you're not just being intellectually lazy and responding negatively to the unknown.

Why did you leave a tenure-track job?

new__username

Because my job at RStudio is basically the same as a tenured position except:

1. I don't have to write grants.
2. I don't have to go to meetings.
3. I only teach when I want to.

That said, I do miss working with awesome students as much as I used to.

I use ggplot2 constantly, so thank you so much for creating and maintaining such wonderful software. I have three very different questions:

1. **About two years ago, Douglas Bates (primary author and maintainer of the R package lme4 for mixed effects modeling, for those unaware) announced that he was quitting R to focus on development in Julia due to perceived problems with the R language and packaging requirements. I see that you briefly commented on the mailing list back then in response. I'm not a developer, but I trust Douglas Bates if he says that there are problems. Do you agree with him, and if so, do you think that there has been any progress in correcting these issues, particularly with regard to CRAN?**

2. **I just peeked at your Instagram and saw some cocktails with Cynar, Ramazzotti, and Amaro Nonino. Do you have a favorite amaro?**

3. **A friend of mine, who is a graduate student who spends her summers at Rocky Mountain Biological Laboratory in Colorado, wanted me to tell you that the grad students there this summer held a "ggplotluck," in which they all made food based on your family recipes. I hope you find that anecdote amusing and not creepy.**

zeurydice

1. I'm hopeful that the R consortium is going to help resolve some of the problems with the package development process because it's going to be able to apply significant funding to the problems. There's nothing public yet, but I'm confident that there will be significant improvement in the next 6-18 months.

2. I don't have a favourite amari yet. I like them all - the bitterer the better!

3. That is awesome! It is less creepy than the person who wanted to give a signed head shot of me to her husband for their wedding ;)

Is there any good reason to use SAS or SPSS these days?

dashfjd

You have a whole bunch of money you want to get rid of? 😜

How can you get soooooooo much done? It is amazing! What are your secrets to productivity? For example, how does a day in your life look like, from waking up to bedtime?

music05

Most of my practical tips are in my quora answer, but here's a bit more about my typical day.

I normally wake up somewhere between 6 and 7. I try to immediately spend an hour writing - in an ideal world I do that before I check twitter and email, but that doesn't always happen. Depending on whether I'm currently involved in more writing or programming heavy projects, I spend the next few hours programming or writing. I go to yoga at 12-1, and then eat lunch. I spend the rest of the afternoon (until 6) doing more writing/programming.

the WINNOWER

I'M HADLEY WICKHAM, CHIEF SCIENTIST AT RSTUDIO AND CREATOR OF LOTS OF R PACKAGES (INCL. GGPLOT2, DPLYR, AND DEVTOOLS). I LOVE R, DATA ANALYSIS/SCIENCE, VISUALISATION: ASK ME ANYTHING! : REDDIT

On Fridays, I make a significant effort to get to inbox zero, and to handle my other responsibilities (reviewing papers, misc pull requests etc). I try to ignore email as much as possible during the rest of the week. I also try and schedule random meetings on Friday as much as possible.

I avoid working on the weekends/

Hi Hadley, thank you for everything you've done and continue to do! An emerging theme in the R community that seemed prevalent at useR this year was developing the next generation of interactive visualization. Since you've worked a lot in this area and have obviously thought about it a lot, where do you think things are going and when do you think we'll get to a point where things are reasonably mature, as static plotting is now? Do you think we'll have one major solution, like plotting via the web browser with something like D3 as a low level driver, or do you forsee multiple solutions akin to base/lattice/ggplot?

gravity

I'm putting my time behind ggvis which will eventually play a similar role to ggplot2, except that the grammar will also extend to interactivity. Unfortunately I haven't spent as much time on it as I'd like (because I got distracted by all the data import packages) but I'm hoping to spend a big chunk of 2016 on it.

That said, I think one of the reasons that ggplot2 was successful is that it only need to handle the most common 90% of visualisation. You could alway use another R package if ggplot2 didn't do exactly what you wanted. I see htmlwidgets playing a similar role for ggvis. There are lot of awesome existing special purpose js libraries, and it's easy to create R bindings for them with htmlwidgets.

Shifting topics a bit: from your experience at Rice University, would you like to go back to teaching? And would you recommend the MSc Statistics to someone pursuing a Data Analyst track? Cheers.

RobFP

I enjoy the act of teaching, but I don't enjoy a lot of the infrastructure around it. For example, in most classes you can not assume that students will be self-motivated about your topic, and you can not assume most students actually know how to learn a new topic. That means you need to provide a lot of scaffolding to make sure students do what's in their best interests. To me, a big thing is assigning weekly homeworks, because it forces people to work through their knowledge of a new subject while it's still fresh. But that obviously adds a lot of infrastructure - you need to make sure grading is fair, provides useful feedback, and timely.

For any masters, I think you need to be ruthless about evaluating it from an investment point of view. What are you going to get out of it? What is it going to cost (in both time and money)? You need to find out what typically graduates from a MSc project go on to do. Given the current massive demand for data scientists, I'd be very concerned about a MSc project where the majority of students didn't go on to jobs earning $100k+.

Hi Hadley,

Thanks for coming here, and being a massive positive force in the R landscape.

I teach students their first steps into stats and R. I do almost all of my own plotting inggplot and a lot of data transformation with dplyr as they produce such clean code and I feel I work a lot faster with them. But I still teach in base R.

Would you say one should first learn base before switching to these packages, or would I serve

the WINNOWER

I'M HADLEY WICKHAM, CHIEF SCIENTIST AT RSTUDIO AND CREATOR OF LOTS OF R PACKAGES (INCL. GGPLOT2, DPLYR, AND DEVTOOLS). I LOVE R, DATA ANALYSIS/SCIENCE, VISUALISATION: ASK ME ANYTHING! : **REDDIT**

my students better by getting (some of these) packages into their repertoire as early as possible?

Ax3m4n

I'm obviously biased, but I do think your students will be more effective if you teach them ggplot2 and dplyr early on. I don't think the code itself is key, but having code that connects with powerful ways of thinking about a problem is really important. dplyr and ggplot2 (and tidyr and ...) give you small building blocks that you can flexibly recombine to solve new problems. I think that makes it easier for students to learn (because the individual pieces are consistent) and to apply their knowledge to new scenarios (because they can recombine the pieces in new ways).

Many of my packages (especially stringr and lubridate) were developed because I was frustrated by teaching the idiosyncrasies of base R.

David Robinson also has some nice thoughts on the issue.

To what extent do you believe R will displace proprietary paid software (such as SAS) in the private/corporate world over, say, the next 5-10 years?

TedMcGriff

It's hard to tell, but it feels like the writing is on the wall for SAS. Lately I've been talking to an number of companies who are switching from SAS to R primarily because college grads now know R and not SAS.

(That said, SAS is a very profitable company and statistics is only a small part of what they do. I'm sure they'll be around for decades yet)

(I've also heard rumours that SAS uses R internally for rapid prototyping, and is training more of its employees in R)

Thanks for your many contributions in the world of R. I noticed that you used to program in Java, which was my first main language and has led to me quite like the structure of S4 classes in R. I wondered if there are any situations in which you would advocate the use of S4 over S3?

LHMarsh

I'm not a big fan of S4 because while I think it's good for solving problems where you have a complex network of classes and methods, those aren't the sort of problems that generally crop up when using R.

Hi there! I am a huge fan of RStudio and use it every day. I recently upgraded to the newest version and love it. In particular, the new functionality in View is great.

Lately I've been working with a lot of XML and JSON encoded datasets, and the result is a bunch of really complicated lists that are hard to view or parse in any convenient way. Do you know of any tools that would make this easier? There are times when it would be ideal to look at them in some sort of tree view, and other times when a tabbed viewer would be better. Any thoughts?

another30yovirgin

No, but we know about the problem and hope to tackle it in the next 6-12 months.

I'm fortunate enough to have worked my way into a job in data analysis and statistics but unfortunately don't any formal background beyond an undergraduate into-to-stats class. I've

done OK learning things as I go, often through blogs and YouTube. Problem is, I don't know what I don't know, and can only learn about things insofar as I know what to type in a search bar query. And even then, even relatively simple formulas are often presented literally in Greek and rarely broken down into layperson-friendly terms. Nate Silver's book The Signal and The Noise was pretty good at introducing higher-level analytical approaches and presenting applicable case studies. Can you recommend any other books or sources that do a good job introducing statistial and data analysis concepts in an easy-to-read, layperson-friendly approach?

[TedMcGriff](#)
I don't have a good recommendation, unfortunately. I'd love to find one, so hopefully someone else will suggest a book that they found helpful.

Hi Hadley,

You're quite a role model to me not because of your excellent work, but because of how approachable and helpful you are to newcomers of the community. I sincerely appreciate how much you engage with questions on Stack, #rstats, rOpenSci, etc etc (I've met you twice, and you've always been kind and patient with my beginner questions). I'm hopeful you will bring the same spirit of inclusiveness to the the R Consortium.

That said, what sort of content/organizing can someone like myself do to give back to the R community? I really want to be more involved in things, but I'm at a loss at where to start. Roger Peng and Hilary Parker's new [podcast](#) is an excellent idea of content that I'm optimistic for, I really appreciate stuff that blends both the "why" and "how" we use R.

Keep on being you, thanks again

[polisighhh](#)
Thanks for the kind words. If you want to give back, I think writing a blog is a great way. Many of the things that you struggle with will be common problems. Think about how to solve them well and describe your solution to others. The next step is to figure out how to wrap your solutions up into a function and then in a package. Keep your eyes open for similar problems and think about how you can create simple components that can be combined to solve them. This is hard so don't be surprised if it takes a while (years!) to come together.

hey Hadley

I read with some surprise this criticism of you on Dirk Eddelbuettel's (R Foundation member at large, author of the Rcpp package) [blog](#):

Hadley is a popular figure, and rightly so as he successfully introduced many newcomers to the wonders offered by R. His approach strikes some of us old greybeards as wrong---I particularly take exception with some of his writing which frequently portrays a particular approach as both the best and only one. Real programming, I think, is often a little more nuanced and aware of tradeoffs which need to be balanced. As a book on another language once popularized: "There is more than one way to do things."

This is a particularly unctuous criticism, since it purports to speak on behalf of others, but setting that aside: has this been a recurrent problem for you (in working in open source software development where prestige and acclaim take the role of financial windfall), being criticized by those jealous of your prominence?

Do senior R figures appreciate and recognize that you've completely reworked R into a peerless

data munging tool? You haven't popularized R--you've made it orders of magnitude more intuitive and powerful.

you_miami

I do get some criticism from Dirk because I promote a certain approach to doing things and I don't tend to talk about the other approaches. This is not unreasonable criticism, and because of my prominence in the community, I do need to be careful about how I position my work relative to others. But generally, my work is aimed at new comers to R and people who are experts in other fields - I don't want to confuse them with a deep discussion of the nuances and alternative approaches. Instead I want to focus on a single way of doing things that I think is most effective for the most number of people.

Hi Hadley,

What are your thoughts on online data science courses from sites like Coursera and Udacity? Do you have any recommendations (tips/advice/books/courses etc) for anyone interested in getting into the Data Science field? Do you ever use other programming languages like Python in your data science work?

riraito

Unfortunately I haven't taken any online data science classes, so I can't give any good advice. I think it's worthwhile to try a few to see what works for you, and then pick one and stick with it even if the going gets tough (you don't want to waste too much time switching between different programs)

Can you remember a time where the use of statistics dramatically changed your opinion on something? A scenario where the stats disproved many of your preconceived notions about a topic?

rhiever

No, but I try to conscientiously update my beliefs about things based on research. This seems to be depressingly uncommon behaviour, even amongst academics.

Hello Hadley, my question is: how would you feel if R is slowly turning into a corporate commercial software (like SAS for example) and many people now are beginning to look for alternatives in pythons or Julia? Any effort to keep the Open source R a live thing?

wajdix

I don't think R is turning into anything like SAS. The vast majority of development effort is open source (e.g. most of what we do at RStudio is open source), and the commercialisation I think is currently only helping the R community.

I think the R consortium is a great example of how increased commercialisation of R helps the whole community: big companies give back to the community to help improve R for everyone.

If R did not exist, what would be your language of choice for data analysis/visualization?

exxplicit

If R didn't exist when I first started writing statistics code, probably ruby, because I was gung ho about web development in rails.

If I had to start over now, probably either python or julia. Or maybe javascript.

Hadley, I'm an undergrad attempting to build my first R package (one with the goal of expanding data sonification capabilities in RStudio). Do you have any advice/ recommendations for this endevour. Many thanks!

perfettiful1
Read http://r-pkgs.had.co.nz. If you live on the west coast, try and come to my R course - we have an 80% discount for students.

As a current PhD student... You don't have tenure...? If you don't, is there any hope for any of us? Or was this a personal decision based on what R was offering you?

MrLegilimens
I left Rice before the tenure process. Since then I've had tenured offers, but I love my job at RStudio!

Hadley, thanks for doing an AMA. Big fan here and I have two questions:

1) Do you ever sleep? I attended a workshop taught by you (which was excellent). I noticed that while teaching the workshop you were also able to answer questions on the ggplot forum, update your notes and code, and tweet all at the same time. How do you do it?

2) Do you miss academia at all?

florence_craye
1. Yes, I sleep a lot :) I do seem to be naturally quite good at multitasking, although to get stuff done, I find that I really have to focus on one thing at a time.

2. No.

Do you think traditional methods of statistical inference will become less popular or diminished by increased computational power and advanced techniques in data science? Or will they always be closely intertwined?

BeastHotel
I think they'll remain closely intertwined.

I've read that you're very opposed to in-place operations, at least fordplyr. Can you elaborate as to why you feel strongly about that? Would dplyr's speed be more comparable to data.table if you did implement that?

buckhenderson
Pure functions are much easier to reason about and I just don't care about (computer) performance that much.

Thank you for the wonderfulness that is ggplot.

When I publish manuscripts, I always include the R script that contains the quantitative analysis and code for figures, the goal is to make the work I do as reproducible as possible. However, there was a change when opts() changed to theme(), and broke all my code appended to ggplot objets. This meant that someone would have to debug the code submitted with the paper. I now add version numbers for all packages alongside R in the text of the manuscript, but this is a bit cumbersome. As functions are altered or taken away, it risks breaking the existing scripts, yet we are not allowed to update published material.

My question to you is this - how do we keep our data reproducible if R packages change over time?

Thalesian

I think you have to track package versions in a machine readable way, along with tools to readily install older versions on other computers. See packrat for one approach for this.

Why is there no (good) GUI for R?

Developing a solid GUI for SAS has helped millions of people who would never have bothered to learn its syntax to be come statistically literate. I like R as is, but analyses could be easier to learn, more reproducible, and no less powerful if the program had a high quality (i.e. more developed than RStudio or RCommander) point-and-click interface.

What is the reason the R community feels is not necessary?

Zaungast

I think there's no good GUI for R because in some sense a GUI would be inimical to the spirit of R. R is all about giving you freedom to do whatever you can imagine (even if it's a bad idea); a GUI is all about restricting your options to keep you in a safe space.

That said, people are still working on GUIs, particularly for teaching. R commander is an older approach, intRo is a modern web based (shiny) approach.

What is the best way for a programming newbie to learn R?

AmericanResearch

I've heard good things about the data science certificate on coursera.org.

I'm working R for data science with Garrett Grolemund, but that won't be ready for 6-12 months.

I've used ggplot from the early days, and have witnessed its huge impact on how R is used for data visualisation. Is there a "philosophy" or reasoning behind it that has made it so intuitive and flexible?

Thanks for all the great tools you've given us over the years!

nailface

It was heavily influenced by The grammar of graphics and the idea of fluent interfaces

Why aren't there more practical guides on data analysis? I would like to see something along the lines of 'these are the typical problems/choices you face at this stage, this is how most researchers handle this'. Or even a FAQ style thing. At the minute, it seems to be 'here's the theory/maths, here's examples' instead of the other way round. One option could be a point and clicky 'typical analyses' interface in R that fully explains everything. I realise there's a danger in this but I think that there might be solutions to this.

reflexdoctor

I'm working on R for data science with Garrett Grolemund. The JHU group has also been writing a lot of books lately.

Thanks for all the work you do to improve R. It's a privilege to be able to ask the famous Hadley Wickham a question.

THE WINNOWER

I'M HADLEY WICKHAM, CHIEF SCIENTIST AT RSTUDIO AND CREATOR OF LOTS OF R PACKAGES (INCL. GGPLOT2, DPLYR, AND DEVTOOLS). I LOVE R, DATA ANALYSIS/SCIENCE, VISUALISATION: ASK ME ANYTHING! : **REDDIT**

1. **Other than ggplot2, which is your favourite package that you've developed? Or if that's too difficult, the package that you use the most?**
2. **What is R's biggest weakness? Speed/massive datasets/something else?**
3. **What do you think will be the biggest changes to come to R over the coming five years?**
quitelargeballs

1. Hmmmm, I really like stringr because it's simple. It's easy to get your head around and it solves a real problem. I also really like xml2, because it makes working with xml so much less painful

2. I think the biggest weakness is that R doesn't have a cadre of full time professional programmers working to make it better. R core is all volunteers who have other full time jobs.

3. I think a big challenge is going to be fighting the big data hype. Not every problem needs big data to solve it, and no big data system will be able to match R's fluidity and power for in-memory data.

Hi Hadley, could you please update us on what's coming next for your visualisation libraries? I'd love to try ggvis properly, but currently use ggplot2 faceting heavily, so am remaining there for now.

Cheers for the amazing productivity boost you've given us!

pobsprogramme
Lots and lots and lots of work on ggvis :/

I picked up R in grad school. In hindsight, I'm amazed that it and programming in general was not more prevalent in my undergrad classes or even high school.

A lot of teachers and students are not even aware that R exists as an option, so where do you see programming education in 5 years? 10 years? What is RStudio's roll?

Thank you for your time and ggplot2 in particular!

StephenHolzman
I hope programming education improves. I think there is a vast audience of people who could benefit from programming, but current approaches turn them away.

I hope that RStudio is able to put more time and effort into better learning experiences. It's definitely something we talk a lot about, but at this point in time the main challenge is ensuring that RStudio is a viable (profitable) company so that we're around for the long term.

How does the rise of mobile devices and small-screen viewing impact the Grammar of Graphics, and the output of tools like ggplot2 and ggvis? Do there need to be extra optimizations made to ensure the readability of fonts/points/etc?

minimaxir
Yes, probably. But I think the biggest challenges arise when you start dealing with interactive/streaming/dynamic graphics. You need to make sure that enough of the plot stays the same so that you can make effective comparisons over time, while still being able to show new data.

Kansas City or Memphis BBQ?

Bigtuna546
Kansas City

the WINNOWER

I'M HADLEY WICKHAM, CHIEF SCIENTIST AT RSTUDIO AND CREATOR OF LOTS OF R PACKAGES (INCL. GGPLOT2, DPLYR, AND DEVTOOLS). I LOVE R, DATA ANALYSIS/SCIENCE, VISUALISATION: ASK ME ANYTHING! : **REDDIT**

Hi Hadley,

Will there be a hadleyverse in Julia? You have a very clear vision of how things should be done (and its working great in R); there's a lot of promise for Julia but they need guidance on interfaces for data analysis. Would you help guide them?

I wish for the 'move the computation to the data' paradigm to be consistent across R and Julia. I think you are THE person to be driving that vision forward.

patshipan

I'm happy to talk to people in the Julia community. But personally, in order to be effective, I have to maintain a ruthless focus, and I love R and want to make it better.

A few questions related to your impressive productivity:

- **Do you sleep? If so, how much?**

- **What motivates you, gives you the inspiration and energy to stay so productive?**

- **Many of us humble mouth-breathing commoners suffer from what is commonly referred to as "procrastination." Do you have any experience with this? If so, what's your approach to avoiding it?**

Thanks!

smsessio

1. About 8 hours a night.

2. It's hard to say. I'm definitely motivated by the feeling of a job done right - even it's code that probably no one will ever look at, I enjoy the feeling of having done the right thing. I do also enjoy all the positive feedback from people who have found my work helpful.

3. I used to procrastinate quite a lot. Three things I found helpful:

    1. Feeling good: the new mood therapy. I found this book really helpful at understand how emotions work, and how you can adjust your thinking to make them help you, and not harm you.
    2. Your elusive creative genius: I really like this framing of genius, as something that comes to you. You can't force it, you just have to be open to it.
    3. Structured procrastination - keep procrastinating, but turn it into something useful!

Thanks for all your great work for R! I don't what I'd do without *roxygen2* **and** *testthat***!...**

Is it at times difficult working for/with RStudio, given a company needs to make money but R (and its Foundation) is essentially a non-profit? For example, I'm an "emacs weirdo" and so don't use RStudio; do you find it hard to work on projects at RStudio while still keeping support for other parts of the R ecosystem?

willpearse

No, because my role at RStudio is to make R awesome ☺ I have no direct responsibility to make money for RStudio, except that if R is more useful, more people will use it, and so more people will use RStudio and then more people will buy our commercial products.

What's your opinion on apache spark and it's R integration?

ExorXmas

Still too early to say. It's very young, but shows a lot of promise.

Redditor for 9 years, but 27 comment karma?!!. What gives??

BobBeaney
Basically a lurker - I skim https://www.reddit.com/r/programming/, but rarely post

Thanks for all your awesome work with RStudio!

I'm having difficulty creating a program that models pursuit curves given input to certain variables. I have my equation, and RStudio creates a great static plot.

Maybe you can help point me in the right direction for "animating" this plot over time? Drawing the curve slowing, like a pirate ship chasing a merchant ship. =)

RhapC
I'd start by looking at shiny and the animation package.

I love RStudio and use your packages every day. Is there a rule of thumb or a secret for developing a new R package? I've got a series of tools I use for processing a particular type of data and was considering putting it together in a package. How specific should you go with your tools or how much do you leave up to the user (data filtering or cleaning for example)? When are you going to give an advanced R lecture in the southeast? Thanks!

Spamicles
I think it's important for a package to be cohesive - the functions should be aimed at a common problem. I don't think packages need to be big, and indeed are often better if they're kept small. Read Gabor Csardi paean to micropackages at https://github.com/metacran/ropensci-commcall-2015/blob/master/commcalls-6.pdf

What's the best way to thank you for your efforts? Also, do companies hire you to make private packages?

is_it_fun
Send me an email? I love to hear specific examples of how my work has helped you.

I don't do any private consulting at the moment.

hey hadley

silly little nitpicky question: the fantastic, life changing magrittr package and the fantastic, life changing tidyr package have a name conflict in theextract function (so too does the development version of lme4, but that's a different question.)

any plans to address this?

you_miami
Not in the short term. I don't usually load magrittr, so it doesn't affect me 😊