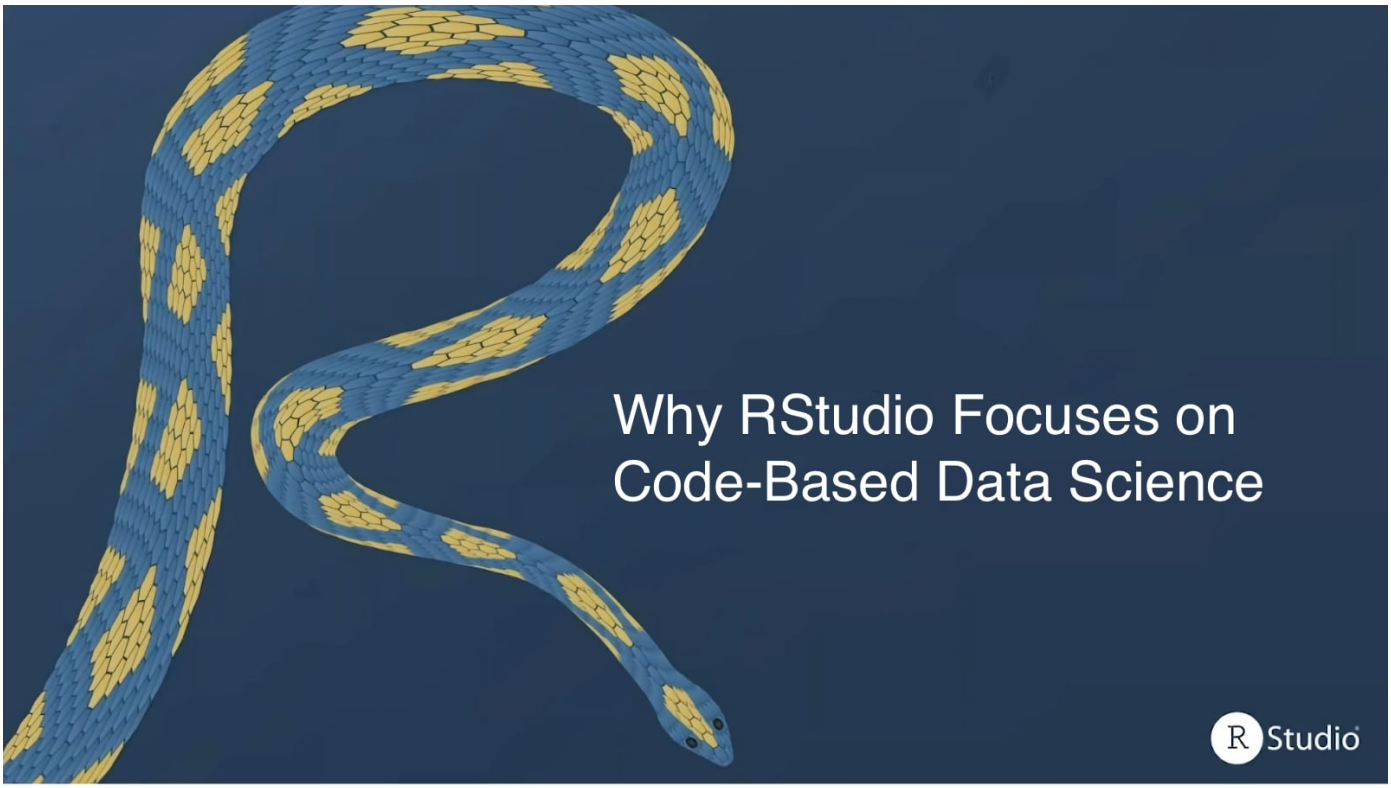
Why RStudio Focuses on Code-Based Data Science

Lou Bajuk, Carl Howe

2020-11-17

Categories: Data Science Leadership Tags: Connect RStudio Package Manager RStudio Server Pro



Michael Lippis of The Outlook podcast recently interviewed RStudio's Lou Bajuk to discuss data science with R and Python, and why RStudio encourages its customers to adopt a multi-lingual data science approach. During the interview, Michael and Lou examined three main topics:

1. RStudio's mission to support open source data science
2. How and why RStudio supports R and Python within its products
3. How business leaders are delivering value from data science investments

I've extracted the most interesting parts of the podcast interview below and edited the quotes for clarity and length. You can listen to the entire interview here.

RStudio's Mission To Benefit Open Source Data Science

Mike: What has been the focus of RStudio since its inception?

Lou: From the beginning, our primary purpose has been to create free and open source software for data science, scientific research, and technical communication. We do this because free and open source software enhances the production and consumption of knowledge and really facilitates collaboration and reproducible research, not only in science, but in education and industry as well. To support this, we spend over half of our engineering resources developing free and open-source software as well as providing extensive support to the open-source data science community.

Mike: How does RStudio help organizations make sense of data regardless of their ability to pay?

Lou: We do this as part of our primary mission around supporting open source data science software. It allows anyone with access to a computer to participate freely in a global economy that really rewards and demands data literacy. So the core of our offerings which enables everyone to do data science is, and will always be, free and open source.

However for those organizations that want to take the data science that they do in R and Python and deploy it at scale, our professional products provide an enterprise-ready modular platform to help them do that. This platform addresses the security, scalability, and other enterprise requirements organizations need to allow their team to deploy their work, collaborate within their team, and communicate with the decision makers that they ultimately support.

Mike: So what is RStudio's public benefit?

Lou: We announced in January that we're now registered as a Public Benefit Corporation (PBC) in Delaware. We believe that corporations should strive to fulfill a public beneficial purpose and that they should be run for the benefit of all of our stakeholders. And this is something that's really a critical part of our founder and CEO JJ Allaire's philosophy.

Our stated public benefit is to create open source software for scientific and technical computing, which means that the open source mission we've been talking about is codified into our corporate charter. And as a PBC, we are committed to considering the needs of not only our shareholders, but all our stakeholders including our community, our customers and employees.

And as part of this, we're now also a Certified B Corporation®, which means we've met the certification requirements set out by the nonprofit B Lab®. That means that we've met the

highest verified standards of things like social and environmental performance, transparency, and accountability.

Multilingual Data Science

The interview continued with Lou diving into why RStudio has committed to supporting both R and Python within its products.

Mike: Why are R and Python in RStudio, and what challenges are you addressing?

Lou: In talking to our many customers and others in the data science field, we've seen that many data science teams today are bilingual, leveraging both R and Python in their work. And while both languages have unique strengths, these teams frequently struggle to use them together.

So for example, a data scientist might find themself constantly needing to switch contexts between multiple development environments. The leader of a data science team might be wrestling with how to share results from their team consistently, so they could deliver value to the larger organizations while promoting collaboration between the R and Python users on their team. The Dev Ops and IT admins spend time and resources attempting to maintain, manage, and scale separate environments for R and Python in a cost effective way.

To help data science teams and the organizations they're in solve these challenges, and in line with our ongoing mission to support the open source data science ecosystem, we've focused our professional products on providing a single centralized infrastructure for bilingual teams using R and Python.

Mike: Is it possible for a data scientist to use R and Python in a single project?

Lou: Absolutely. And there multiple ways that can be done. One of the most popular is an open source package called `reticulate` that we've developed. `Reticulate` is an open source package that is available to anyone using R. It provides a comprehensive set of tools for interoperability between Python and R, including things like:

- Calling Python from R in a variety of ways, whether you're doing something with R Markdown, importing Python modules, or using Python interactively within an R session.
- Translating data objects between R and Python
- Binding to different versions of Python, including virtual and Conda environments.

Mike: What about data scientists whose primary language is Python? What does RStudio provide for them?

Lou: First off, we've been working on making the RStudio IDE a better environment for Python coding. In addition to the `reticulate` package we just discussed, we've just announced some new features in the upcoming release of our IDE, RStudio 1.4. This includes displaying Python objects in the environment pane, viewing Python data frames, and tools for configuring Python versions and different Conda virtual environments. All this is going to make life easier for someone who wants to code Python within the RStudio IDE.

Secondly, for a team where you might have multiple different data scientists who have different preferences for what IDEs they want to use, our pro platform provides a centralized workbench supporting multiple different development environments. In addition to our own IDE we support Jupyter notebooks and Jupyter Lab as development environments, and we're working on more options for the near future. This includes Visual Studio Code, which we're going to be announcing a beta of very shortly.

And finally with our platform, Python-oriented data scientists can create data products and interactive web applications such as Plotly, Streamlit, or Bokeh in their framework of choice and then directly share those analyses with their stakeholders.

We believe this ability for Python data scientists to share their results in a single place alongside the data products created by the R data scientists is critical to actually impacting decision making at an organization

Mike: How can data science leaders promote collaboration across a bilingual team?

Lou: Data science leaders often see their teams struggle to collaborate and share work across disparate, open source tools. Often they waste time translating code from one language to another to put it into production. These activities really distract them from their core work. And as a result, their business stakeholders are less likely to see results or must wait longer for them.

With RStudio products, a bilingual team can work together, building off each other's work. Best of all, it can publish, schedule, and email regular updates for interactive analyses and custom reports built in both languages. So you, the data science team, and your stakeholders will always know where to look for these valuable insights.

Mike: Has RStudio done anything to help DevOps engineers and IT administrators deal with the difficulty of maintaining separate environments for each data science language?

Lou: Absolutely. DevOps and IT is a critical stakeholder in the whole process of doing data science effectively in your organization. So with RStudio products, DevOps, and IT can maintain a single infrastructure for provisioning, scaling and managing environments for both R and Python users. This means that IT only needs to configure, maintain and secure a single system.

A single system also makes it easy for IT to leverage their existing automation tools and other analytic investments and provide data scientists with transparent access to their servers or Kubernetes or SLURM clusters, directly from the development tools those data scientists prefer. They can easily configure all the critical capabilities around access, monitoring, and environment management. And of course RStudio's Support, Customer Success, and Solutions Engineering teams are here to help and advise these teams as they scale out their applications.

Mike: How do business stakeholders view bilingual data science teams?

Lou: Ultimately most decision makers really don't care what language a data science insight was created in. They just want to be able to trust the information and use it to make the right decision. That's why we're so focused on making it easy for data scientists to create these data products, regardless of whether they're R or Python, and then easily share them with their different stakeholders.

Delivering Value from Data Science Investments

Mike and Lou wrapped up by discussing how businesses can improve the value they derive from their data science.

Mike: What would you say to business leaders that are worried about the value of their data science investment?

Lou: One of the big challenges that organizations face with data science is not just how they solve today's problems, but how they ensure that they continue to deliver value over time. Too many organizations find themselves either struggling to maintain the value of their legacy systems, reinventing the wheel year after year, or being held over a barrel by vendor lock-in.

To address this, we recommend a few approaches:

One is to build your analyses with code, not clicks. Data science teams should use a code-oriented approach because code can be developed and adapted to solve similar problems in the future. This reusable and extensible code then becomes the core intellectual property for an organization. It'll make it easier over time to solve new problems and to increase the aggregate value of your data science work. This is why code-first data science is really a critical part of RStudio's philosophy and our roadmap.

The second major approach is to manage your data science environments for reproducibility. Organizations need a way to reproduce reports and dashboards as projects, tools, and dependencies change. You'll often hear about repeatability and reproducibility when talking about a heavily regulated environment like pharmaceuticals, and it's certainly particularly

critical there. However, it's critical for every industry; otherwise your team may spend far too much time attempting to recreate old results. Worse, you may get different answers to the same questions at different points in time, which really undermines your team's credibility.

And third, deploy tools and interactive applications to keep insights up to date, because no one wants to make a decision based on old data. Publishing your insights on web-based interactive tools such as the RStudio Connect platform helps keep your business stakeholders up-to-date and gives them on demand access and scheduled updates. By deploying insights in this way, your data scientists are free to spend their time solving new problems rather than solving the same problem again and again.

Mike: Has RStudio done anything to empower business stakeholders with better decision-making?

Lou: This is really a key focus of ours. Many data science vendors out there focus on creating models and then putting these models into "production". which typically means integrating these models into some system for automated decision-making. For example, a model might determine what marketing offer to present to someone who visits a website.

Though our products certainly support this through the ability to deploy R and Python models as APIs to plug into other systems, our focus is broader. We want to make it easy for a data science team to create tailored reports, dashboards, and interactive web-based applications, using frameworks like Shiny that they can then easily and iteratively share with their decision makers. This iterative and interactive aspect is critical because decision-makers will invariably come back with questions like "What if you run this analysis on a different time period?" or "What if this parameter is different?".

Interactive applications give these decision-makers tremendous flexibility to answer their own "what if?" questions. When it's easy for the data scientist to create a new version, tweak the code, and redeploy it, it's also more convenient for the decision maker. It allows them to get a timely answer that's really super focused on what they actually need as opposed to a generic report.

We call these reports tailored or curated because of their flexibility. Open source data science means that these teams can provide their stakeholders with exactly the information they need in the best format for presenting that information rather than being constrained by the black box limitations of a BI reporting tool.

Mike: Can you provide the Outlook series audience with an overview of RStudio Team?

Lou: RStudio Team is a bundle of our professional software for data analysis, package management, and data product sharing. The Team product is a way of getting all three products, but each of

these products can also be purchased individually to fit into and complement an organization's existing data science investments.

The first component is RStudio Server Pro, which provides a centralized work bench for analyzing data and then developing and sharing new data products and interactive applications. This is the platform where the data scientists develop their insights.

Secondly, RStudio Connect is a centralized portal for distributing these dashboards, reports and applications created by the data scientists, whether they're written in R and Python. This includes the ability to schedule and send email reports to your community of users and to provide all the access control and scalability and reproducibility that a modern enterprise really needs.

Thirdly, RStudio Package Manager supports both the development side (RStudio Server Pro) and the deployment side (RStudio Connect) by managing the wealth of open source data science packages you might need to create and run these analyses. Open source data science hosts a world of people creating these great packages on the cutting edge of statistics and data science but managing these packages over time can be very difficult. RStudio Package Manager makes maintenance and reproducibility much easier.

Mike: All right, Lou. So, can you share a use case with our audience?

Lou: We have a ton of great customer stories at rstudio.com, but one of my favorites is Redfin. Redfin is a technology-powered real estate brokerage that serves more than 90 metropolitan areas across the U.S. and Canada. Now when Redfin was smaller, they used to do a lot of planning using basic data models implemented in spreadsheets and gathering input from emails or files saved in Google drive.

But Redfin wanted to get better, smarter answers. They wanted to make these models more complex and to scale these models to handle the increasing scope of the business. And they found that spreadsheets just wouldn't work anymore. They weren't able to apply the more statistical approaches for forecasting that they wanted and maintaining the formulas and spreadsheets was error prone and slow. Plus, the amount of time that it took to consolidate user input into these spreadsheets limited how many iterations of their models they could run. These workbooks then would be painfully slow, sometimes taking 10 more minutes to open up and use. Sometimes they would crash, leaving people unable to use them at all.

Redfin used RStudio products to move their data models from spreadsheets to a much more reproducible and scalable data science environment. They saw our products as a way to replicate the interactivity that users loved in spreadsheets, but host all this on a server that was easy to access and maintain. This approach allowed them to build in all those complex statistical approaches that they wanted while still keeping the end interface simple for the end users.

Mike: All right, Lou. Where can the audience get more information on RStudio's solutions?

Lou: All the information is available on our website at rstudio.com and there we talk about our products. We also make it easy to either download our products and try them out, or set up a call with our great sales team to help provide some guidance and answer any questions you have. I also encourage your listeners to follow the RStudio blog at blog.rstudio.com, where we write about many of the themes I talked about today as well as share updates on our products and our company.

For More Information

If you'd like to learn more about some of the topics discussed in this interview, we recommend exploring:

- An overview of how RStudio helps multi-lingual data science teams at R & Python: A Love Story .
- RStudio's mission and status as a Public Benefit Corporation.
- More examples of the problems RStudio's customers are solving with our products.

← How California Uses Shiny in Production to Fight COVID-19 ⊹ RStudio 1.4 Preview: New Features in RStudio Server Pro →

Subscribe

News & Events

Upcoming webinars ↪