

R is Hot

How Did a Statistical Programming Language Invented in New Zealand Become a Global Sensation?


By David Smith

Much in the same way that social networking, reality TV and craft beer were considered marginal fads before gaining widespread acceptance from the mainstream culture, the fast-growing popularity of R strongly suggests that it is heading toward a similar level of acceptance by the analytic community.

R has already won praise and plaudits from established media outlets such as the New York Times, Forbes, Intelligent Enterprise, InfoWorld and The Register. When you consider that R is a high-level computer programming language designed mostly for *quants* (the nickname for a subspecies of geeks who focus on quantitative analysis), the adoring media attention seems nothing short of astounding.


So it's entirely fair to ask: Why all the hoopla? Why is an esoteric programming language created in the early 1990s by two academics in New Zealand suddenly all the rage? Why is R so hot?

Let's examine some of the reasons behind the rising popularity of R. As is the case with almost every new trend, there are underlying economic and social factors – nothing just “happens,” there are always root causes. For example, it's no secret that our digital information systems generate new data at an unimaginably fast pace – sometimes it seems as though we're drowning in data.



"There are very few things that SAS or SPSS will do that R cannot, while R can do a wide range of things that the others cannot."

Robert A. Muenchen
Author, *R for SAS and SPSS Users*



Despite this apparently inexhaustible supply of new data, the perceived value of data is rising, which has led to the development of quicker, better and more powerful methods for analyzing complex sets of numbers. The current generation of analytic solutions are cumbersome and costly, however, which has opened the door to newer and less expensive techniques for crunching big numbers. Many of these newer and less costly techniques are written in R, which has rapidly become the “common language” of people whose careers or livelihoods are driven by data.

“R is the most powerful and flexible statistical programming language in the world,” says Norman Nie, a nationally recognized scholar in the fields of survey research, quantitative social science and political behavior. A co-founder of SPSS in the late 1960s, Nie is currently CEO and president of Revolution Analytics, a company based in Palo Alto that provides commercialized versions of R programs.

“What was once a secret of drug-development statisticians at pharmaceutical companies, quants on Wall Street, and PhD-level statistical researchers around the globe (not to mention pioneers at Web 2.0 companies like Google and Facebook) is suddenly becoming mainstream,” says Nie.

Robert A. Muenchen, the author of *R for SAS and SPSS Users*, writes that R has already had a profound impact on research in a variety of fields that rely on quantitative analysis to generate usable information.

Since its release in 1996, R has dramatically changed the landscape of research software. There are very few things that SAS or SPSS will do that R cannot, while R can do a wide range of things that the others cannot. Given that R is free and the others quite expensive, R is definitely worth investigating.

"I can't think of any programming language that has such an incredible community of users. If you have a question, you can get it answered quickly by leaders in the field. That means very little downtime."

Mike King
Quantitative Analyst,
Bank of America

More Than a Programming Language

Unlike traditional analytic software products, R is a fully-fledged programming language. But R has already evolved into more than just a language – R represents a radically different approach to the challenges posed by increasingly large and complex sets of data. In that respect, it is something of a cultural phenomenon.

R is an open source project, which means that it depends on a worldwide community of active developers to grow and evolve. Like Linux, the most famous open source project, R isn't "owned" by any single person or entity. R is maintained and supported by thousands of individuals who use it and who contribute to its ongoing development.

The members of this global community serve as R's parents and custodians – and they take their responsibilities seriously. Like doting parents, they take pride in the achievements of their offspring – and they are quick to leap in when they perceive a problem.

"I can't think of any programming language that has such an incredible community of users," says Mike King, a quantitative analyst at Bank of America. King uses R to write programs for capital adequacy modeling, decision systems design and predictive analytics. "If you have a question, you can get it answered quickly by leaders in the field. That means very little downtime."

"The R core group is an extremely talented collection of statisticians who have a great vision and who really think about what they're doing."


Abhijit Dasgupta
Consulting
Biostatistician,
National Institutes of Health

Critical Mass and Going Viral

R was created in 1993 by Ross Ihaka and Robert Gentleman at the University of Auckland in New Zealand. It's called R for the simple reason that both of its creators have first names beginning with the letter "R." Some believe that R's single-letter name represents a sort of homage to the S language, since R is an open-source descendent of S and much of the code written for S runs unaltered in R. The S language was developed by a Bell Labs team that included John Chambers, for which he won the prestigious ACM Software System award in 1998.

Ihaka and Gentleman had set out to create a language that would make it easier for them to teach their introductory data analysis courses. But news of the new language spread quickly, and in 1995 they were convinced to make the R source code available under the terms of the Free Software Foundation's GNU General Public License. Their decision to share R freely was a seminal moment in the annals of analytic software development.

As interest in R surged, a core group of dedicated volunteers coalesced around the project. This core group of leading statisticians and computer scientists from around the world is now the project's official leadership team. They are the guardians of the R language, and oversee changes



"R is becoming ubiquitous, so if you're starting a huge project and you don't have a lot of programmers ... you go to colleges and hire people who can be productive when they walk in the door. That's a huge benefit."

Robert Sudol

Sr. Development
Manager in Fixed
Income Technology,
AllianceBernstein


and implementations of new features to the R source code on a six-month cycle. They also provide guidance, support and advice to R users through a very active mailing list.

"R got lucky because the core group is an extremely talented collection of statisticians who have a great vision and who really think about what they're doing," says Abhijit Dasgupta, a consulting biostatistician at the National Institutes of Health. "It also built up a very active user community around this core group. As a result, the customer support for R through various online forums is amazingly good. That's one of the truly great benefits of R – it has a fantastic user community that keeps growing. It's infectious."

According to Muenchen, who has actually measured the popularity of data analysis software through the careful analysis of Internet traffic, "R is the most discussed software by roughly a two-to-one margin, followed by Stata then SAS."


R's skyrocketing popularity translates into more than just bragging rights – the R user community is now so large that it generates new R programs (called "packages") at an astonishing pace. It's almost as if the R community has achieved critical mass and been transformed into one gigantic, self-organizing virtual factory that produces new R software with clockwork regularity.

How does this self-organizing virtual software factory work? Let's look at one more or less typical member of the R user community, Glenn Meyers.



Meyers is a vice president of research at ISO Innovative Analytics. He holds a bachelor's degree in mathematics and physics from Alma College in Alma, Michigan, a master's degree in mathematics from Oakland University, and a Ph.D. in mathematics from the State University of New York at Albany. He is a Fellow of the Casualty Actuarial Society and a member of the American Academy of Actuaries. If you're a casualty actuary, you are probably already familiar with Meyers. He's won many of the top actuarial prizes and awards, he gives speeches and he sits on international committees.

Meyers also writes a regular column for *Actuarial Review*. When he writes about a new technique for analyzing data, he often includes the code for the new method in his column. Most of his code is written in R, because R has become the common language – the *lingua franca* – of statistical analysis.



"New methods show up in R before they show up in other packages. R is definitely on the cutting edge."

Michael Elashoff


Director of
Biostatistics,
CardioDX

"You put out your R code and it becomes immediately usable," says Meyers. "People download the code and just start using it."


Large commercial software vendors will rarely develop new programs unless there's a large enough market to justify their development costs. And it can take years for large vendors to bring new programs to market.

The R community, on the other hand, develops and releases new software continuously, thanks to the contributions of thousands of people like Meyers. "The most powerful reason for using R is the community," he says.

Or as Robert Sudol of AllianceBernstein puts it, "the more people who use R, the more powerful it becomes."



Sudol uses R to predict future economic trends by analyzing seasonal data and looking for patterns or anomalies. It's a complicated job that requires creative thinking and improvisation. "If



"You put out your R code and it becomes immediately usable."

Glenn Meyers
Vice President of
Research, ISO
Innovative Analytics

you're trying to do something that's not in the code set, you go out and find an R package ... and then you snap it right in and start using it," Sudol explains. "There are so many people out there making modifications and enhancements, you're going to find something you can use."

Sudol says that he sees "a lot of parallels between R and Linux. R is becoming ubiquitous, so if you're starting a huge project and you don't have a lot of programmers ... you go to colleges and hire people who can be productive when they walk in the door. That's a huge benefit."

R also offers benefits to companies that are trying to reduce the amount of money they spend on renewing licenses with traditional enterprise software vendors such as SAS and SPSS. "The nicest thing about R is that it's free," says Sudol. Choosing an R package over a traditional software product can literally save you "hundreds of thousands of dollars," says Sudol.




Power from Elegance

If the R movement has a genuine rock star, it's probably Hadley Wickham. He's an assistant professor and the Dobelman Family Junior Chair in Statistics at Rice University. He's written and contributed to more than 20 R packages, and he's won the John Chambers Award for Statistical Computing.

Most of Wickham's research focuses on making data analysis better, faster and easier. He is especially interested in using visualization techniques to improve how data and models are understood. In other words, he's all about making it easy to use R.

"R was designed from the ground up to deal with common data problems," says Wickham. "Compared to other programming languages, it's designed to help you do the kinds of things that you do most often when you're performing data analysis. For example, R has data frames built into the core language. It's such a natural structure, and it makes working with data much easier. But very few other languages have data frames built in."



"R was designed from the ground up to deal with common data problems."

Hadley Wickham
Author, *ggplot2:
Elegant Graphics for
Data Analysis*.

Because R was created *by* statisticians *for* statisticians, it's already loaded with many of the crucial features required to accomplish the everyday tasks of statistical analysis. The very design of the R language is often described as "elegant" – in other words, R is in tune with the way statisticians think and work.

For example, says Wickham, "In statistics, it's really critical to keep track of missing values. That's when you don't know what a value is, but you need some way of indicating it. R keeps track of that for you, so that if you add a number to a missing number, you still don't know what that number is and R will keep track of it. That's important."



No Need to Reinvent the Wheel

Precisely because R is a programming language – as opposed to being a pre-fabricated piece of software – new analytic techniques that are written in R can be saved and re-used. So when R users discover something fresh and exciting, they have two options that are not generally available to users of pre-fab software:

1. They can share the new techniques with other R users, inside their organizations and all over the world.
2. They can reproduce and re-use the new techniques they have discovered.

These are not trivial or minor advantages – they represent enormous potential value. The ability to save and re-use improvised functions means that you’re not forced to reinvent the wheel each time that you run an analytic operation. Try doing that in SAS or SPSS and you’re in for a long haul.

The ability to share new R code through forums hosted by CRAN (Comprehensive R Archive Network) and other groups ensures a state of continuous evolution. Bluntly put, the world of R never sits still.

“New methods show up in R before they show up in other packages,” says Michael Elashoff of CardioDX, a molecular diagnostics company that collects data from multiple sources and builds predictive models in R that help physicians detect cardiovascular diseases in their patients.

“We do a lot of predictive model development on complex data sets, so the ability use and evaluate new statistical methods is important to us. Especially in the last couple of years, many of these newer methods have been showing up as R packages first. R is definitely on the cutting edge,” says Elashoff.

Zubin Dowlaty, VP / Head of Innovation & Development at Mu Sigma, has a similar take on the value of R. Headquartered in Chicago, Mu Sigma is a global analytics services company providing business decision support services to clients in data-intensive industries such as pharma, insurance, financial services, CPG/retail, healthcare and technology. All of that means that Mu Sigma is in the business of analyzing data – big time.

“The large ecosystem of statisticians all over the world adding new functions and packages to the R system is a huge benefit,” says Dowlaty. “State-of-the-art algorithms are available quickly through the R platform.”

The R platform has become so comprehensive that it now represents a “one-stop shop” for analytical techniques, says Dowlaty. “Most of the techniques you need to drive analytics into the business are available through R – everything from statistical to machine learning and optimization techniques. Unlike other vendors, like SAS or SPSS, R provides everything in one go-round.”

High Quality Graphics, Made Easy

R is especially useful for generating charts and graphics, quickly and easily. The ability to create visual plots of complex data is more than just a handy trick; it’s an incredibly important step in the analysis of data because it enables you to literally “see” the patterns and anomalies hidden within the data.

The New York Times has been a leader in the use of charts and graphics that make it easier for readers to get the gist of complicated stories. Amanda Cox, a graphics editor at the *Times*, says R is particularly valuable in deadline situations when data is scant and time is precious. “If you can picture it in your head, chances are good that you can make it work in R,” says Cox. “R makes it easy to read data, generate lines and points, and place them where you want them. It’s very

“Most of the techniques you need to drive analytics into the business are available through R – everything from statistical to machine learning and optimization techniques. Unlike other vendors, like SAS or SPSS, R provides everything in one go-round.”

Zubin Dowlaty

VP / Head of
Innovation &
Development, Mu
Sigma

“It’s very flexible and super quick. When you’ve only got two or three hours until deadline, R can be brilliant.”

Amanda Cox

Graphics Editor, New
York Times

"R helps us show our clients how they can improve their processes and effectiveness by enabling our consultants to conduct analyses efficiently.

John Lucker
Consulting Principal -
Advanced Analytics
and Modeling,
Deloitte Consulting,
LLC

"I've found R
incredibly useful for
processing data
quickly."

Peter Aldhous
San Francisco Bureau
Chief, *New Scientist*
magazine

flexible and super quick. When you've only got two or three hours until deadline, R can be brilliant."

When Michael Jackson died in 2009, the *Times* quickly prepared a graphic timeline showing how the artist's songs had performed on the Billboard Hot 100 chart from 1971 to the present. It would have been difficult or impossible to prepare a similar chart on deadline using other analytic techniques.

Peter Aldhous, the San Francisco bureau chief of *New Scientist* magazine, has used R to generate information that is subsequently used by graphic designers to create some of the charts that illustrate his articles. But he also uses R to generate simple plots that allow him to perceive quickly what's really going on underneath the data he collects. For a journalist, the ability to draw quick insights from data is absolutely invaluable.

"I've got a Ph.D. in animal behavior, so I have some statistical training," says Aldhous. "R is great for doing exploratory work that gives me an idea of what the distributions look like. I've found it incredibly useful for processing data quickly."

Recently, Aldhous investigated complaints about certain academic papers on stem cell research being subjected to "obstructive" reviews, resulting in delays or spurious rejections by peer-reviewed journals. Using an R package to generate a quick series of box plots and scatter plots, he saw that papers from scientists outside the US seemed to take longer to get accepted and published. He was then able to follow up and analyze the data in R, using the most appropriate statistical and graphical methods: Cox proportional hazards regression and Kaplan-Meier curves.

In an article headlined "Hey, Green Spender," Aldhous and colleague Phil McKenna examined the gap between consumer perception and environmental realities across multiple industries such as retail, media, travel and leisure, food and beverages, technology, construction and chemicals.


When the data was plotted, the differences between the perceptions and the realities were immediately visible – and the reporters knew they were on the right track.

"It's not just about producing graphics for publication," Aldhous explains. "It's about playing around and making a bunch of graphics that help you explore your data. This kind of graphical analysis is a really useful way to help you understand what you're dealing with, because if you can't see it, you can't really understand it. But when you start graphing it out, you can really see what you've got."

R makes it possible for people who aren't professional analysts to create high quality charts and graphics such as maps, 3-D surfaces, image plots, scatter plots, histograms, bar plots and pie charts.


"R is far from easy when you first encounter it, especially if you're not a programmer who is used to working in the command line. It has a very steep initial learning curve," says Aldhous. "But once you get to grips with its conventions and quirks, and if you study the documentation, then it becomes easy to plug different variables into the same code to create a series of related graphics."

When Aldhous hit a snag or got in over his head, he reached out to the R community for help. "The community is delightful and incredibly helpful," he says. "I could not have done all of this without expert help."



"R is the most powerful and flexible statistical programming language in the world."

Norman Nie
CEO, Revolution Analytics



Building a Business

The value of R to business is borne out by the experiences of John Lucker and his team of advanced analytics professionals at Deloitte Consulting LLP. John is a Deloitte Consulting Principal and leads the firm's Advanced Analytics and Modeling (AAM) practice, one of the leading analytics groups in the professional services industry.

When the group was launched fourteen years ago, its main focus was solving vexing business problems for clients in the insurance industry. One of the challenges facing the industry was the lack of robust analytic processes for supporting critical underwriting decisions. This challenge was particularly acute in the rapidly growing commercial insurance industry. "The commercial insurance underwriting process was rigorous but also quite subjective and based on intuition," says John.

Convincing experienced underwriting management to change their time-honored traditions was not an easy task. That's where R proved exceptionally helpful in recent years. "R enables us to communicate our analytic results in appealing and innovative ways to non-technical audiences through rapid development lifecycles," says John. "Sometimes people know they have a problem, but they don't know how to fix it. And sometimes they don't even know they have a problem. R helps us show our clients how they can improve their processes and effectiveness by enabling our consultants to conduct analyses efficiently."

The group's success with clients in the insurance industry became a blueprint for expanding into new markets. Today, the Deloitte Consulting Advanced Analytics and Modeling practice also serves clients in healthcare, banking and financial services, retail, consumer products, telecomm, automotive, media, hospitality, public/state/federal sector and other major industries. R played an important role in growing the practice by allowing it to offer robust analytics addressing the specific needs of clients in a variety of markets.

"I find the diversity of R solutions and add-ons very appealing," says John. "R has served as a catalyst in the marketplace. It forces everyone to raise their game, and it incents software developers to enhance their offerings. Everybody becomes more competitive and users of analytic tools benefit."

Changing, Transforming and Evolving

With thousands of contributors and two million users worldwide, R is a truly global phenomenon. Unlike traditional commercial software for data analysis, R is both flexible and extensible. Supported by an active community of users and developers, R is constantly changing to meet the changing needs of our rapidly shifting global economy.

The popularity of R is no fluke or fad. R has become the common language of data analysis because it was designed – from the ground up – as a practical system for handling the real-world challenges of complex data sets. R-based programs are applied routinely to solve problems in real-time trading, finance, risk assessment, forecasting, biotechnology, drug development, social networking and more.

But the wide acceptance of R as the *lingua franca* of statistics is based on its unique ability to change, to transform and to evolve. When new techniques in statistical analysis are discovered, they tend to emerge as R packages first – years before those innovations are incorporated in traditional enterprise software products.

Thanks to its open-source roots, R has spread virally across the map. It has become both ubiquitous and indispensable. The R community supports development, innovation and continuous improvement. New players are welcome and encouraged. The R eco-system has become a fertile breeding ground for novel ideas and original ways of thinking about numbers.

No one can foretell the future of quantitative analytics, but it's safe to wager that a good deal of it will be written in R.

About David Smith

David is the Vice President of Marketing at Revolution Analytics, the leading commercial provider of software and support for the open source R statistical computing language. David is the co-author, with Bill Venables, of the official R manual *An Introduction to R*. He is also the editor of [Revolutions](http://blog.revolutionanalytics.com) (<http://blog.revolutionanalytics.com>), the leading blog focused on “R” language, and one of the originating developers of ESS: Emacs Speaks Statistics. You can follow David on Twitter as [@revodavid](https://twitter.com/revodavid)

About Revolution Analytics

Revolution Analytics was founded in 2007 to foster the R community, as well as support the growing needs of commercial users. Our name derives from combining the letter "R" with the word "evolution." It speaks to the ongoing development of the R language from an open-source academic research tool into commercial applications for industrial use.

Through our [Revolution R products](#), we aim to make the power of predictive analytics accessible to every type of user & budget. We provide free and premium software and services that bring high-performance, productivity and ease-of-use to R – enabling statisticians and scientists to derive greater meaning from large sets of critical data in record time.

We also offer our full-featured production-grade software to the academic community for [FREE](#), in order to support the continued spread of R's popularity to the next generation of analysts.

For customers such as Pfizer, Novartis, Yale Cancer Center, Bank of America and others, our flagship [Revolution R Enterprise](#) product stands for faster drug development, reduced time of data analysis, and more powerful and efficient financial models.

Please visit us at www.revolutionanalytics.com

Reproducing this Document

Copyright 2010 Revolution Analytics. Some rights reserved: you may share, unmodified, this text or excerpts of this text provided you attribute it to Revolution Analytics and link to this URL: www.revolutionanalytics.com/R-is-Hot/. Licensed under a [Creative Commons Attribution-NoDerivs 3.0 Unported License](#).