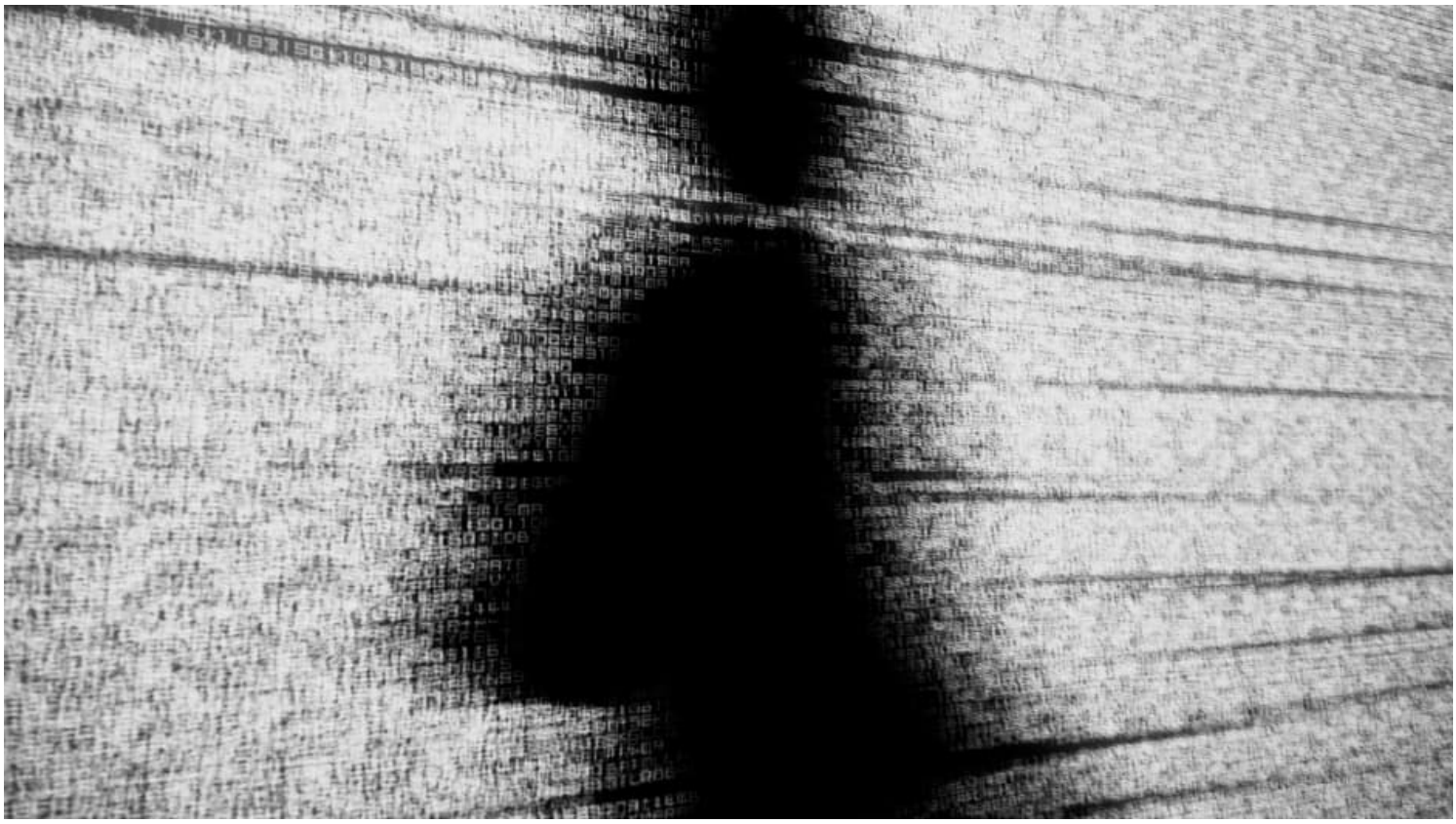
Here's a roadmap to the latest and greatest tools in data science, and when you should use them.



[IMAGE: FLICKR USER R2HOX]

BY ANNA NICOLAOU

6 MINUTE READ

This story contains interviews with Michael Driscoll, CEO of Metamarkets; Paul Butler, data scientist at Chango and formerly at Facebook; and Niall O'Connor, vice president at Bank of America.

The big data frenzy continues. It's permeating nearly every industry, flooding companies with more and more information, and making software dinosaurs such as Excel look more and more inept. Data crunching is no longer just for nerds, and the need for sophisticated analysis and powerful, real-time processing is greater than ever.

So what are the best tools to sift through gigantic data sets? We talked to data hackers about their favorite languages and tool kits for hardcore data analysis.

R

It would be downright negligent to start this list with any language other than R. It has been kicking around since 1997 as a free alternative to pricey statistical software, such as Matlab or SAS.

But over the past few years, it's become the golden child of data science–now a household name not only among nerdy statisticians, but also Wall Street traders, biologists, and Silicon Valley developers. Companies as diverse as Google, Facebook, Bank of America, and the *New York Times* all use R, as its commercial utility continues to spread.

R has simple and obvious appeal. Through R, you can sift through complex data sets, manipulate data through sophisticated modeling functions, and create sleek graphics to represent the numbers, in just a few lines of code. It's likened to a hyperactive version of Excel.

R's greatest asset is the vibrant ecosystem has developed around it: The R community is constantly adding new packages and features to its already rich function sets. It's estimated that more than 2 million people use R, and a recent poll showed that R is by far the most popular language in data science, used by 61% of respondents (followed by Python, with 39%).

It's also catching on on Wall Street. Traditionally, banking analysts would pore over Excel files late into the night, but now R is increasingly being used for financial modeling, particularly as a visualization tool, says Niall O'Connor, vice president at Bank of America. "R makes our mundane tables stand out," he says.

R is maturing into its role as a go-to language for data modeling, although it's power becomes limited when a company needs to produce large-scale products, and some say it's already being usurped by other languages.

"R is more about sketching, and not building," says Michael Driscoll, CEO of Metamarkets. "You won't find R at the core of Google's page rank or Facebook's friend suggestion algorithms. Engineers will prototype in R, then hand off the model to be written in Java or Python."

Paul Butler famously used R to build a Facebook map of the world back in 2010, proving the rich visualization capabilities of the language. He doesn't use R as often as he used to, though.

"R is becoming a bit passé in industry, because it's slow and clunky with large data sets," said Butler.

So what is he using instead? Read on.

PYTHON

If R is a neurotic, loveable geek, Python is its easygoing, flexible cousin. Python is rapidly gaining mainstream appeal as a hybrid of R's fast, sophisticated data mining capability, and a more practical language to build products. Python is intuitive and easier to learn than R, and it's ecosystem has grown dramatically in recent years, making it more capable of the statistical analysis previously reserved for R.

"It's the big one people in the industry are moving toward. Over the past two years, there's been a noticeable shift away from R and towards Python," says Butler.

In data processing, there's often a trade-off between scale and sophistication, and Python has emerged as a compromise. IPython notebook and NumPy can be used as a scratchpad for lighter work, while Python is a powerful tool for medium-scale data processing. Python also has the advantage of a rich data community, offering vast amounts of toolkits and features.

Bank of America uses Python to build new products and interfaces within the bank's infrastructure, but also to crunch financial data."Python is broad and flexible, so people flock to it," says O'Donnell.

Still, it's not the highest-performance language, and only occasionally can it power large-scale, core infrastructures, says Driscoll.

JULIA

The vast majority of data science today is conducted through R, Python, Java, MatLab, and SAS. But there's still gaps to be filled, and Julia is one newcomer to watch.

Julia is still too arcane for widespread industry adoption. But data hackers get giddy when talking about its potential to oust R and Python from their thrones. Julia is a high-level, insanely fast and expressive language. It's faster than R, and potentially even more scaleable than Python, and fairly easy to learn.

"It's up and coming. Eventually, you'll be able to do anything you could have done in R and Python, in Julia," says Butler.

Youth is holding Julia back, for now. The Julia data community is in its early stages, and more packages and tools are needed before it can viably compete with R or Python.

"It's young, but it's gaining steam and very promising," says Driscoll.

JAVA

Java, and Java-based frameworks, are found deep in the skeletons of the biggest Silicon Valley tech companies. "If you look inside Twitter, Linkedin, or Facebook, you will find that Java is the foundational language for all of their data engineering infrastructures," says Driscoll.

Java doesn't provide the same quality of visualizations R and Python do, and it isn't the best for statistical modeling. But if you are moving past prototyping and need to build large systems, Java is often your best bet.

HADOOP AND HIVE

A flock of Java-based tools have popped up to meet the enormous demand for data processing. Hadoop has exploded as the go-to Java-based framework for batch processing. Hadoop is slower than some other processing tools, but it's insanely accurate and widely used for backend analysis. It pairs nicely with Hive, a query-based framework that runs on top.

SCALA

Scala is another Java-based language and, similar to Java, it's increasingly becoming the tool for anyone doing machine learning at large scales, or building high-level algorithms. It's expressive, and also capable of building robust systems.

"Java is like building in steel. Scala is like working with clay that you can then put into a kiln and turn into steel," Driscoll says.

KAFKA AND STORM

What about when you need rapid, real-time analytics? Kafka is your friend. It's been around for five years, but just recently became a popular framework for stream processing.

Kafka, which was born inside of Linkedin, is an ultra-fast query messaging system. The downside to Kafka? It's too fast. Operating in real time lends itself to error, and occasionally Kafka misses things.

"There's a trade-off between precision and speed," says Driscoll. "So all the big tech companies in the Valley use two pipelines: Kafka or Storm for real-time processing, and then Hadoop for batch processing system that will be slow but super-accurate."

Storm is another framework written in Scala, and it's gaining enormous traction for stream processing in Silicon Valley. It was acquired into Twitter which, unsurprisingly, has a huge interest in rapid event processing.

Honorable mentions:

MATLAB

MatLab has been around for eternity, and despite its price tag, it's still widely used in very specific niches: research-intensive machine learning, signal processing, and image recognition, to name a few.

OCTAVE

Octave is very similar to MatLab, except it's free. Still, it's rarely seen outside of academic signal processing circles.

GO

GO is another newcomer that's gaining steam. It was developed by Google, loosely derives from C, and is gaining ground against rivals such as Java and Python for building robust infrastructures.

You Might Also Like:

Eight Strategies For Tackling Legacy Code You Didn't Write

The Founder Of Duolingo Tells Us The Secret To Creating Value From Chaos

Why Does The World Need More Programming Languages?

VIDEO

I Took An Improv Class To Conquer My Fear Of Public …

Writer Katharine Schwab is on a personal mission to become a better version of herself. To…

I Took An Improv Class To Conquer My Fear Of Public Speaking

NOW PLAYING

I Took An Improv Class To Conquer My Fear Of Public Speaking

Kate Hudson On How Her Mother's Trailblazing Influenced Her Entrepreneurial Instincts

Trixie And Kat And Dentistry

IDEAS

IDEAS

GoFundMe's New Film Studio Wants To Help Giving Campaigns Go Viral

FAST COMPANY MAGAZINE

SheaMoisture's Tangled Future

IDEAS

How Closing Grocery Stores Perpetuate Food Deserts Long After They're Gone

ENTERTAINMENT

ENTERTAINMENT

How Funny Or Die Plans To Conquer TV Comedy

ENTERTAINMENT

Exclusive: Inside Formula One's Rebranding Strategy

ENTERTAINMENT

How Mother Teresa And Smart Career Advice Inspired "Novitiate" Director Maggie Betts

CO.DESIGN

PRODUCTS

This Artificial Muscle Can Turn Robots Into Superheroes—Or Supervillains

PRODUCTS

20 Gift Ideas For People Who Refuse To Shop On Amazon

UI & UX

The Popular Web Design Tool That's Actually A Privacy Nightmare

FAST COMPANY

SPONSORED CONTENT

This Way Ahead: Preparing For The Future Of Retail

TECHNOLOGY

Where This Supercomputer Is Going, There Are No Hard Drives

FAST COMPANY MAGAZINE

Brain Hacks And Apps To Help You Remember Names, Dates, And More