
What Is a Data Scientist?

By Jonathan Williams

While you may not realize it, you're a data factory, producing data at every turn you take. Grabbing your morning coffee and paying with a credit card? You've produced data for the store and credit card company. Googling directions to your first meeting of the day? You've created data from your search request. Hopping on the bus and swiping your transit card? You just produced data stored with your municipal transportation group. Within a matter of minutes, you've created a handful of data. Multiply this by the rest of the day, the rest of the year, and wow — so much data has been created, just by you.

Around 2.5 quintillion bytes of data are created per day — an enormous sum of data that humankind has never known before. Not only is the volume of the world's data increasing, but it's arriving at a faster and faster pace, with approximately 90% of the world's data having been created within the past two years alone, according to IBM. Not only is data bigger and being created faster; it's also coming in a variety of formats. Just think of the variety of data in your social media behavior alone: You're posting text-based commentary to Twitter, uploading videos to Snapchat, and Instagramming pictures of your pet.

However, the truth is that data by itself doesn't have much value. After all, a pile of numbers and data files is just that: a pile of numbers and data files. The real value of data comes from making sense of the abundance of information. That's why businesses and organizations across countless industries are investing in forward-thinking data talent — to leverage data's predictive power, craft smart business strategies, and drive informed decision making.

DID YOU KNOW?

Data scientist has been named both the No. 1 Job in America and the “sexiest job of the 21st century.”



“Our ability to do great things with data will make a real difference in every aspect of our lives.”

— Jennifer Pahlka, Founder & Executive Director for Code for America

But who is the sharp and strategic person for this job? Enter, the data scientist. Data scientists often serve as a company’s data ambassador, working cross-functionally to provide information across departments and to a broad range of stakeholders.

What makes a data scientist, well, a scientist?

Not all scientists work with beakers and chemicals in the lab. Data scientists, like other scientists, develop questions to answer with data. They investigate unusual patterns, build new models, and are the inquisitive and curious minds behind the data they work with.

A data scientist’s process begins similarly to those in other scientist roles. A central question often motivates their work, such as, how do we fix crowded buses in Chicago?

To answer the question, the data scientist puts on their detective hat and collects data that might be relevant. They could sift through a treasure trove of data from the Chicago Transit Authority, or might consider bringing in information about holidays, weather, or special events happening in the city. The data scientists get down to business and might even refine the original question with a new one: Can the temperature predict bus ridership in Chicago during the hours of 7-10 a.m.?

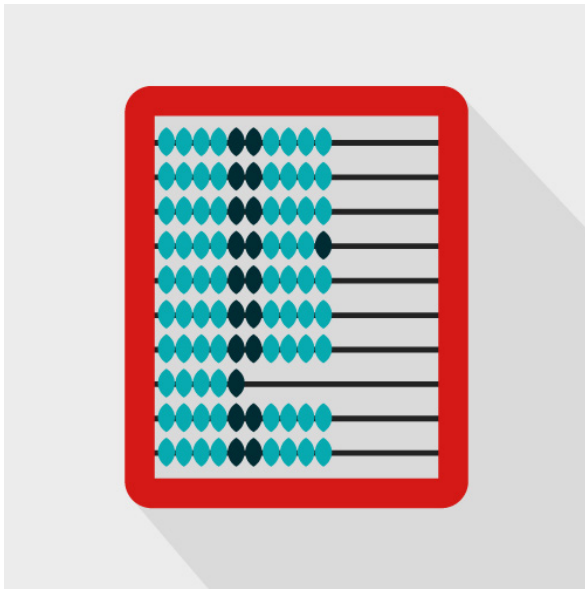
Data scientists are nimble thinkers who ask and refine the right questions, but they also rely heavily on two domains of knowledge that qualify them as scientist: mathematics and programming.

The mathematical knowledge required of a data scientist consists of three main components:

Descriptive statistics. Data scientists are faced with a lot of data, and descriptive statistics — which encompass the ability to describe data — help them understand what’s important.

DID YOU KNOW?

The abacus — developed thousands of years ago — is often referred to as the first computer.



“Data! Data! Data! I can’t make bricks without clay!”

— Sherlock Holmes

Matrix algebra and calculus. Data scientists often work with high-dimensional data (that’s data with lots of variables), which are formatted as matrices (rectangular arrays of data). Matrix algebra and calculus serve as the underpinnings of many of the modeling techniques applied in building inferential statistical models.

Inferential statistics. Data scientists must feel comfortable in a numbers-based environment, while fluidly applying statistical concepts to their work. Just like a chemist knows which chemical to grab from the shelf, a data scientist has to know which statistical model to grab and implement to form a prediction.

Because of the volume of data that data scientists work with, gone are the days of simple tools like the abacus for counting, or a slide rule for multiplication and division. The situation calls for a unique toolset, mainly in the form of a programming language, like scripting languages, SQL and Python, or a more numerical language, like R. These tools can manipulate data and develop predictive models with some work, but take time to learn.

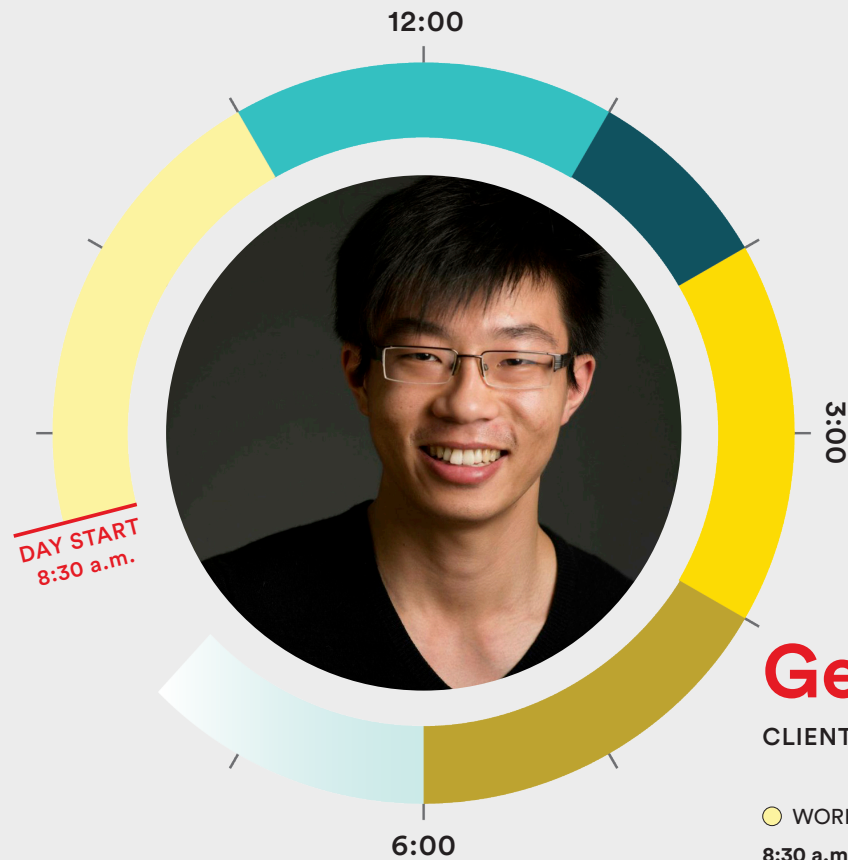
When data scientists mix the right mathematical method with a programming language, they often uncover something unexpected, something the naked eye could never discover in the data.

Where do data scientists work?

If an industry creates data, there’s likely a data scientist behind the scenes working to make use of this data. And, if an industry hasn’t discovered the power of data yet, it will soon!

When we think of data scientists, companies like Amazon and Facebook (which have massive amounts of data) quickly come to mind. Data scientists at these companies might develop recommender systems to suggest relevant products to a customer or develop algorithms to deliver targeted advertisements to users. Thank your data science friends for recommending cute cat socks to you after you purchased cat food on Amazon Pantry.

WHERE DO DATA SCIENTISTS WORK?



George Zhao

CLIENT INSIGHTS ANALYST AT APPNEXUS

WORK WHILE EXERCISING

8:30 a.m.: Morning gym session with my team lead. Between sets we discuss upcoming work and focuses. As a two-person team, it works!

CHECK DATA LOGS

10 a.m.: Check cron logs to ensure that no critical data pipelines failed; they feed important dashboards for our services team.

BUILD FEATURES 'TIL LUNCH

1 p.m.: Build a new feature on an internal flask app to monitor algorithmic performance. Merge and release for the rest of the team! Time for lunch!

DEEP LEARNING PROJECT

2 p.m.: Check in on the deep learning model I'm training as a side project; current test accuracy is disappointing. Note ideas for improvement.

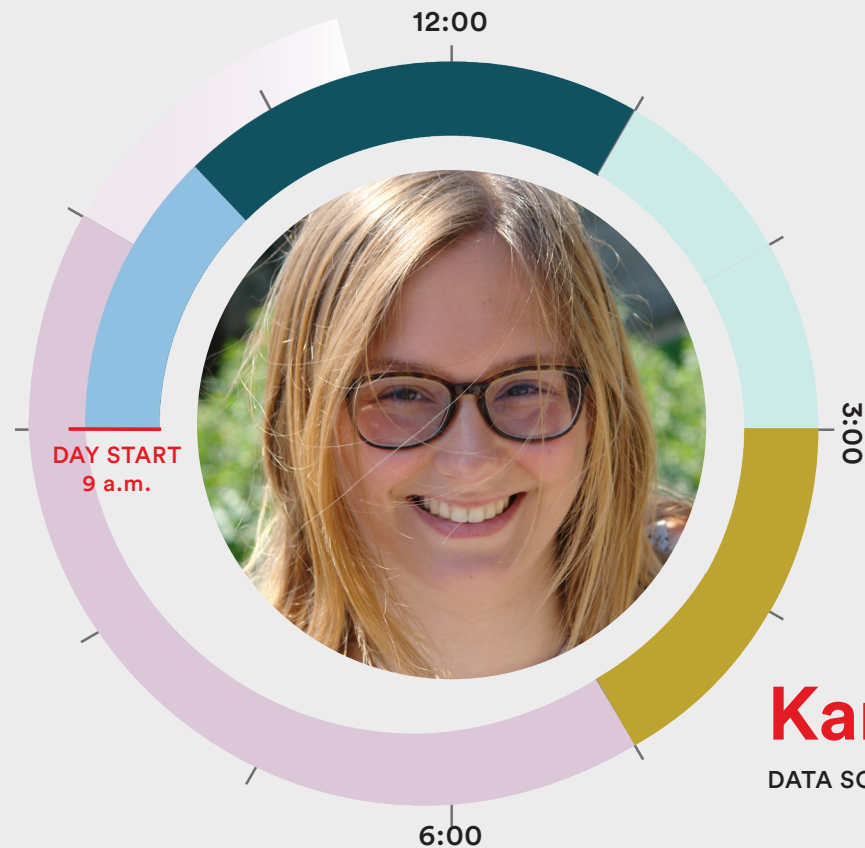
MEET WITH ACCOUNT MANAGERS

4 p.m.: Meet with account management team to scope out an analytics request for a quarterly business review for a large publisher client.

DATA ANALYSIS 'TIL DINNER

6 p.m.: Neck deep in data analysis to investigate a hypothesis around audience segment targeting on our platform. Stop...or get caught up in the excitement and keep going until too hungry.

WHERE DO DATA SCIENTISTS WORK?



Karlijn Willems

DATA SCIENCE JOURNALIST AT DATACAMP

● CHECK NEWS AND DATA TRENDS

9 a.m.: Visit KDnuggets, Hacker News, and DataTau to read up on the latest news in the data science industry; check emails; and reply to questions from DataCamp users.

● CODING

10:30 a.m.: Start writing Python code for a new machine learning tutorial. Machine learning rocks!

○ CHECK AND SCRAPE DATA

1 p.m.: Check the R scraper script I built and scrape textual data. This will be part of an article for the DataCamp community so others can learn from it.

● REVIEW AND FEEDBACK LOOP

3 p.m.: Review guest contributors' code, write down suggestions, and make a pull request (PR) to get my feedback accepted. Plan calls or email them to discuss feedback.

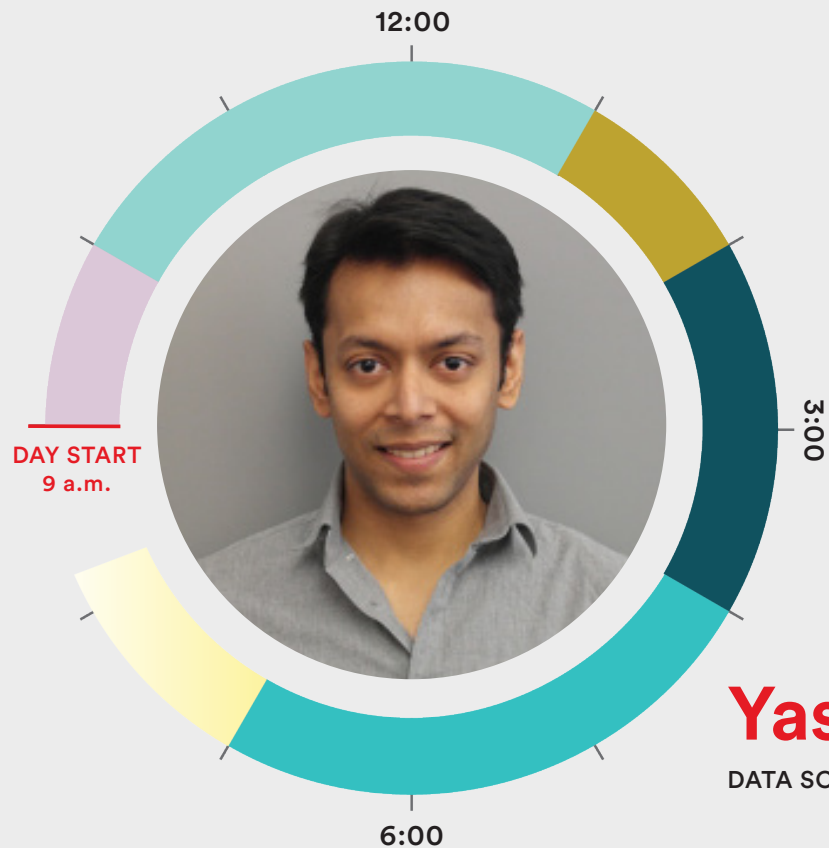
○ NETWORKING + SKILL UP

5 p.m.: Leave for the Brussels Data Science meetup. They're organizing an interesting tutorial series on deep learning with TensorFlow.

○ NETWORKING + BEER

10 p.m.: Connect with peer data science enthusiasts over a good Belgian beer. Check the latest data science news on the train back home.

WHERE DO DATA SCIENTISTS WORK?



Yash Shah

DATA SCIENTIST AT APPNEXUS

● BREAKFAST NETWORKING AND COLLABORATION

9 a.m.: Grab breakfast with a few data science teammates and discuss upcoming projects and ideas, and also help resolve one another's blockers.

● PERFORMANCE CHECK

10 a.m.: Check automated performance snapshot emails and performance metrics for all of yesterday's models.

● PRODUCT MANAGEMENT

1 p.m.: Meet with product manager and discuss the project vision as well as client needs and feedback for the ongoing tests.

● CODE OVER COFFEE

2 p.m.: Grab coffee with the optimization tech engineer to discuss implementation issues and concerns about the code deployed to production.

● PREDICTION MODEL RESEARCH

4 p.m.: Dig deeper into the data to check for trends that could help better inform our prediction model.

● TEAM SPORTS

7 p.m.: End the day with a good workout at the gym or play softball, basketball, or soccer representing our company team.

DID YOU KNOW?

Legendary computer science pioneer Peter Naur coined the term “data science” in 1960.



“All data has its beauty, but not everyone sees it.”

— Damian Mingle, Author and Chief Data Scientist at Intermedix

Data scientists don't only work in technical companies or fields. They can work in the financial industry to develop models for investing or develop computational methods that search for fraudulent transactions. They work in local government to optimize operations or predict the resources used in their city. And, data scientists can even even work for the social good, like those lending their skills through the organization DataKind. For example, a group of DataKind's London-based data scientists teamed up with nonprofit partners to understand what advice individuals seek before becoming homeless. Now, the nonprofit partners have a better sense of advice-seeking patterns associated with a higher chance of homelessness, allowing them to tweak their services and support for this at-risk population.

The joy of a career in data science is that you can combine your chops for data science with another field to find meaningful work. Maybe you're a fashionista with a flair for big data — how about predicting optimal order quantities for next season's collection? The options of data science + you are endless.

What else do I need in my toolbox to be a successful data scientist?

Apart from mathematical and programming skills, successful data scientists often have a strong business acumen or substantive experience in their field. Data scientists help businesses make data-driven decisions. In order to do this, however, they have to understand the underlying business itself. This isn't to say you have to be expert in a specific field or in business in addition to data science, but knowing the business allows you empathize with the needs of the business from day one.

As the person who knows the most about the company's data, effectively communicating that information to others is as important as the number-crunching or programming. A data scientist might bounce between a business's

ABOUT GENERAL ASSEMBLY:

General Assembly teaches in-demand skills for today's tech-driven world. Change your career in our full-time Data Science Immersive course, discover our part-time campus options, or skill up with a mentor on a flexible schedule with our online Data Analysis Circuit. Our courses dive into trendspotting, predictive modeling, and much more to ready you for the fast-growing field of data.

WE TEACH:

- [Full-time Data Science Immersive course >](#)
 - [Part-time Data Analytics course >](#)
 - [Part-time Data Science course>](#)
 - [Part-time online Data Analysis Circuit>](#)
- [Browse our courses at ga.co >](#)

non-technical departments: The marketing team may need help deciphering analytics around a marketing campaign, or the finance department might need to predict next quarter's revenue. The ability to communicate and present technical insights approachably, efficiently, all while speaking to key organizational objects is a differentiator among data scientists.

Don't be just a creator of data; be the part of the data science revolution that shapes how we understand the data around us.

Jonathan Williams is an instructional designer for data education programs at General Assembly.

ga.co