# Advice to aspiring data scientists: start a blog

Last week I shared a thought on Twitter:

> *When you've written the same code 3 times, write a function*
>
> *When you've given the same in-person advice 3 times, write a blog post*
>
> *— David Robinson (@drob) November 9, 2017*

Ironically, this tweet hints at a piece of advice I've given at least 3 *dozen* times, but haven't yet written a post about. I've given this advice to almost every aspiring data scientist who asked me what they should do to find a job: **start a blog, and write about data science**.[1]

What could you write about if you're not yet working as a data scientist? Here are some possible topics (each attached to examples from my own blog):

- Analyses of datasets you find interesting (example, example)
- Intuitive explanations of concepts you've recently mastered (example, example)
- Explorations of how to make a specific piece of code faster (example)
- Announcements about open source projects you've released (example)
- Links to interactive applications you've built (example, example)
- Sharing a writeup of conferences or meetups you've attended (example)
- Expressing opinions about data science or educational practice (example)

In a future post I'd like to share advice about the **how** of data science blogging (such as how to choose a topic, how to structure a post, and how to publish with blogdown). But here I'll focus on the **why**. If you're in the middle of a job search, you're probably very busy (especially if you're currently employed), and blogging is a substantial commitment. So here I'll lay out three reasons that a data science blog is well worth your time.

## Practice analyzing data and communicating about it

If you're hoping to be a data scientist, you're (presumably) not one yet. *A blog is your chance to practice the relevant skills.*

- **Data cleaning**: One of the benefits of working with a variety of datasets is that you learn to take data "as it comes", whether it's in the form of a supplementary file from a journal article or a movie script

- **Statistics**: Working with unfamiliar data lets you put statistical methods into practice, and writing posts that communicate and teach concepts helps build your own understanding
- **Machine learning**: There's a big difference between having used a predictive algorithm once and having used it on a variety of problems, while understanding why you'd choose one over another
- **Visualization**: Having an audience for your graphs encourages you to start polishing them and building your personal style
- **Communication**: You gain experience writing and get practice structuring a data-driven argument. This is probably the most relevant skill that blogging develops since it's hard to practice elsewhere, and it's an essential part of any data science career

I can't emphasize enough how important this kind of practice is. No matter how many Coursera, DataCamp or bootcamp courses you've taken, you still need experience applying those tools to real problems. This isn't unique to data science: whatever you currently do professionally, I'm sure you're better at it now than when you finished taking classes in it.

> *... and that concludes Machine Learning 101. Now, go forth and apply what you've learned to real data! pic.twitter.com/D6wSKgdjeM*
>
> *— ML Hipster (@ML_Hipster) August 19, 2015*

One of the great thrills of a data science blog is that, unlike a course, competition, or job, you can analyze any dataset you like! No one was going to pay me to analyze Love Actually's plot or Hacker News titles. Whatever amuses or interests you, you can find relevant data and write some posts about it.

## Create a portfolio of your work and skills

Graphic designers don't typically get evaluated based on bullet points on their CV or statements in a job interview: they share a portfolio with examples of their work. I think the data science field is shifting in the same direction: the easiest way to evaluate a candidate is to see a few examples of data analyses they've performed.

Blogging is an especially good fit for showing off your skills because, unlike a technical interview, you get to put your "best foot forward." Which of your skills are you proudest of?

- If you're skilled at visualizing data, write some analyses with some attractive and informative graphs ("Here's an interactive visualization of mushroom populations in the United States")
- If you're great at teaching and communicating, write some lessons about statistical concepts ("Here's an intuitive explanation of PCA")
- If you have a knack for fitting machine learning models, blog about some predictive accomplishments ("I was able to determine the breed of a dog from a photo with 95% accuracy")
- If you're an experienced programmer, announce open source projects you've developed and share examples of how they can be used ("With my sparkcsv package, you can load CSV datasets into Spark 10X faster than previous methods")

- If your real expertise is in a specific domain, try focusing on that ("Here's how penguin populations have been declining in the last decade, and why")

Just because you're expecting employers to look at your work doesn't mean it has to be perfect. Generally, when I'm evaluating a candidate, I'm excited to see what they've shared publicly, even if it's not polished or finished. And sharing *anything* is almost always better than sharing nothing.

> *"Things that are still on your computer are approximately useless." -@drob #eUSR #eUSR2017 pic.twitter.com/nS3IBiRHBn*
>
> *— Amelia McNamara (@AmeliaMN) November 3, 2017*

In this post I shared how I got my current job, when a Stack Overflow engineer saw one of my posts and reached out to me. That certainly qualifies as a freak accident. But the more public work you do, the higher the chance of a freak accident like that: of someone noticing your work and pointing you towards a job opportunity, or of someone who's interviewing you having heard of work you've done.

And the purpose of blogging isn't only to advertise yourself to employers. You also get to build a network of colleagues and fellow data scientists, which helps both in finding a job and in your future career. (I've found #rstats users on Twitter to be a particularly terrific community). A great example of someone who succeeded in this strategy is my colleague Julia Silge, who started her excellent blog while she was looking to shift her career into data science, and both got a job and built productive relationships through it.

## Get feedback and evaluation

Suppose you're currently looking for your first job as a data scientist. You've finished all the relevant DataCamp courses, worked your way through some books, and practiced some analyses. But you still don't feel like you're ready, or perhaps your applications and interviews haven't been paying off, and you decide you need a bit more practice. What should you do next?

**What skills could you improve on?** It's hard to tell when you're developing a new set of skills how far along you are, and what you should be learning next. This is one of the challenges of self-driven learning as opposed to working with a teacher or mentor. A blog is one way to get this kind of feedback from others in the field.

This might sound scary, like you could get a flood of criticism that pushes you away from a topic. But in practice, you can usually sense that you're not ready well before you finish a blog post.[2] For instance, even if you're familiar with the basics of random forests, you might discover that you can't achieve the accuracy you'd hoped for on a Kaggle dataset- and you have a chance to hold off on your blog post until you've learned more. What's important is the committment: it's easy to think "I probably could write this if I wanted", but harder to try writing it.

**Which of your skills are more developed, or more important, than you thought you were?** This is the positive side of self-evaluation. Once you've shared some analyses and code, you'll probably find that you were *underrating* yourself in some areas. This affects everyone but it's especially important for graduating Ph.D. students, who spend several years becoming an expert in a specific topic while surrounded by people who are already experts- a recipe for impostor syndrome.

> *Imposter Syndrome: be honest with yourself about what you know and have accomplished & focus less on the difference. pic.twitter.com/VTjS5KdR6Y*
>
> *— David Whittaker (@rundavidrun) April 13, 2015*

For instance, I picked up the principles of empirical Bayes estimation while I was a graduate student, and since it was a simplification of “real” Bayesian analysis I assumed it wasn’t worth talking about. But once I blogged about empirical Bayes, I learned that those posts had a substantial audience, and that there’s a real lack of intuitive explanations for the topic. I ended up expanding the posts into an e-book: most of the material in the book would never qualify for an academic publication, but it was still worth sharing with the wider world.

One question I like to ask of PhD students, and anyone with hard-won but narrow expertise, is “What’s the simplest thing you understand that almost no one outside your field does?” That’s a recipe for a terrific and useful blog post.

## Conclusion

One of the hardest mental barriers to starting a blog is the worry that you’re “shouting into the void”. If you haven’t developed an audience yet, it’s possible almost no one will read your blog posts- so why put work into them?

First, a lot of the benefits I describe above are just as helpful whether you have ten Twitter followers or ten thousand. You can still practice your analysis and writing skills, and point potential employers towards your work. And it helps you get into the habit of sharing work publicly, which will become increasingly relevant as your network grows.

Secondly, this is where people who are already members of the data science community can help. My promise is this: **if you’re early in your career as a data scientist and you start a data-related blog, tweet me a link at @drob and I’ll tweet about your first post** (in fact, the offer’s good for each of your first three posts). Don’t worry if it’s polished or “good enough to share”- just share the first work you find interesting![3] I have a decently-sized audience, and more importantly my followers include a lot of data scientists who are very supportive of beginners and are interested in promoting their work.

Good luck and I’m excited to see what you come up with!

1. It’s also a great idea to blog if you’re *currently* a data scientist! But the reasons are a bit different, and I won’t be exploring them in this post. ↩
2. Even if you do post an analysis with some mistakes or inefficiencies, if you’re part of a welcoming community the comments are likely to trend towards constructive (“Nice post! Have you considered vectorizing that operation?”) rather than toxic (“That’s super slow, dummy!”). In short, if as a beginner you post something that gets nasty comments, **it’s not your fault, it’s the community’s.** ↩
3. A few common-sense exceptions: I wouldn’t share work that’s ethically compromised, such as if it publicizes private data or promotes invidious stereotypes. Another exception is posts that are just blatant advertisements for a product or service. This probably doesn’t apply to you: just don’t actively try to abuse it! ↩

---

## David Robinson

*Data Scientist at Stack Overflow, works in R and Python.*

Email Twitter Github Stack Overflow

### Subscribe

Your email

Subscribe to this blog

### Recommended Blogs

- DataCamp
- R Bloggers
- RStudio Blog
- R4Stats
- Simply Statistics

**Advice to aspiring data scientists: start a blog** was published on November 14, 2017.

YOU MIGHT ALSO ENJOY (VIEW ALL POSTS)