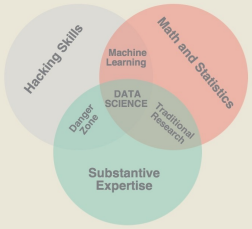
DataCamp
Learn data analysis for free, interactively

Become a

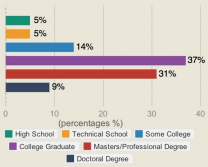
Data Scientist

in 8 easy steps

What's a data scientist?



Typical Background



A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

1 Get good at stats, math and machine learning

Math

> Math Track of Khan Academy
> Linear Algebra by MIT OpenCourseware

Stats

> Intro to Statistics by Udacity
> OpenIntro Statistics

ML

> Machine Learning by Andrew NG (Stanford Online)
> Practical Machine Learning by John Hopkins (Coursera)

2 Learn to code

Computer Science Fundamentals
> CS50x on edX

Grasp end-to-end development
The things you build will be integrated into other systems

Choose a first language
> Open Source: R, Python, etc.
> Commercial: SAS, SPSS, etc.

Learn interactively
> R: DataCamp, try R
> Python: Codecademy, Google Class

3 Understand databases

As a data scientist student, you will often work with data in text files. However, once you enter the industry, a database is almost always used to store data. It's going to be stored in MySQL, Postgres, MongoDB, Cassandra, etc.

Learn more on databases via:

4 Master data munging, visualization and reporting

Data cleaning and munging

WHAT
Data munging is the process of converting one "raw" form into another format for more convenient consumption

TOOLS
> Getting and Cleaning data by John Hopkins (Coursera)
DataWrangler (Beta)
data.table
dplyr

Data visualization

WHAT
Data visualization involves the creation and study of the visual representation of data.

TOOLS
ggvis
D3
vega

Reporting

WHAT
In every data analysis, putting the analysis and the results into a comprehensible report is the final hurdle to take.

TOOLS
tableau
Spotfire (TIBCO Software)
R Markdown

5 Level up with Big Data

When you start operating with data at the scale of the web, the fundamental approach and process of analysis must change. Most data scientists are working on problems that can't be run on single machines. They have large data sets that require distributed processing.

Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.

MapReduce is this programming paradigm that allows for massive scalability across the servers in a Hadoop cluster.

Apache Spark is Hadoop's speedy Swiss Army knife. It is a fast-running data analysis system that provides real-time data processing functions to Hadoop.

6 Get experience, practice and meet fellow data scientists

Practice makes perfect ...

- kaggle — Join in competitions
- meetup — Meet fellow data scientists
- Have a pet project
- Develop your intuition

7 Internship, bootcamp or get a job

The best way to find out whether you are a true data scientist or not is to take the bull by the horns and to enter the real-life jungle of data-analysis and science with your freshly acquired skill set.

Internship	Bootcamp	Job
BEGINNER	INTERMEDIATE	ADVANCED

amazon.com
zipfian

8 Follow and engage with the community

Sites to follow
> DataTau
> Kdnuggets
> fivethirtyeight
> datascience101
> r-bloggers

People to follow
> Hilary Mason
> David Smith
> Nate Silver
> dj patil

Need Data?
quandl

DataCamp
Learn data analysis for free, interactively
@DataCamp_com