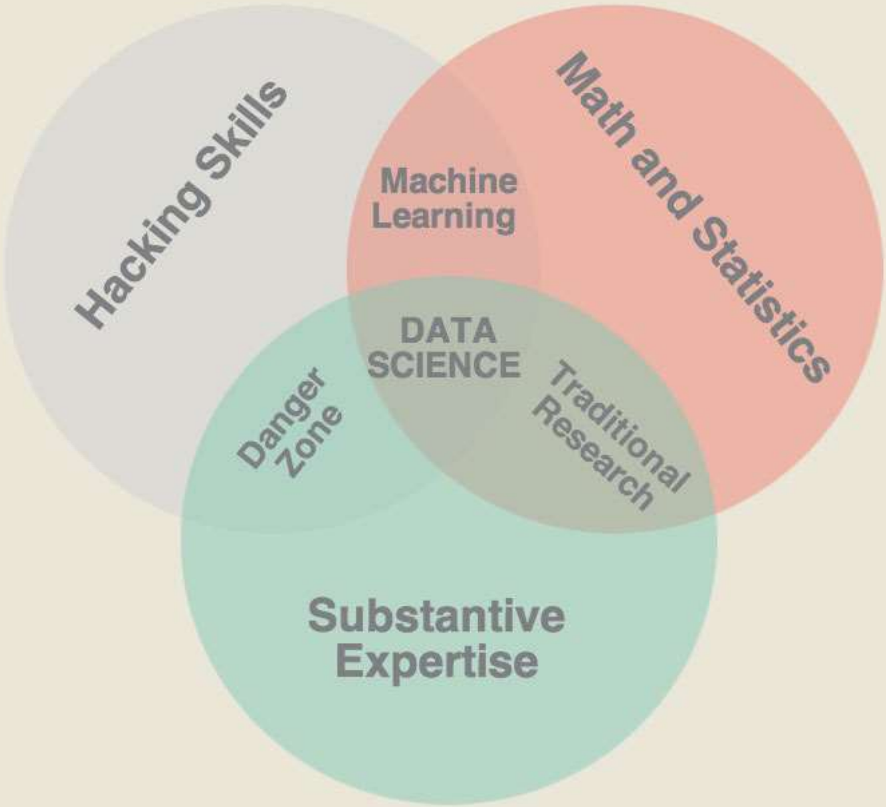
Become a

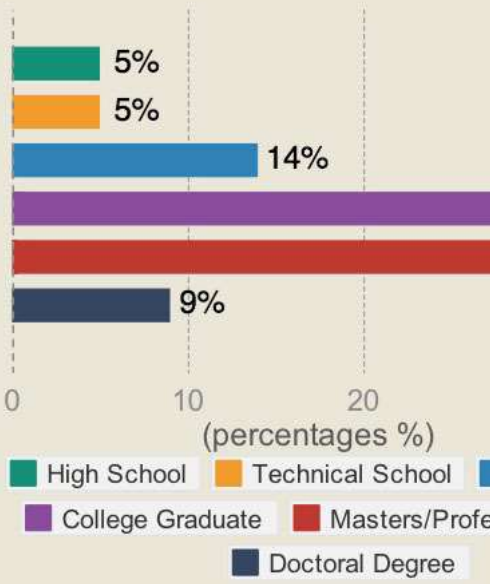
Data Scienti

in 8 easy steps

What's a data scientist?



Typical Background



A data scientist is someone who is better at statistics than any sof
engineer and better at software engineering than any statistici

1 Get good at stats, math and machine learning

2 Learn to code

Math

> Math Track of Khan Academy

> Linear Algebra by MIT OpenCourseware

MIT OCW

Stats

> Intro to Statistics by Udacity

> OpenIntro Statistics

UDACITY

OpenIntro

ML

> Machine Learning by Andrew NG (Stanford Online)

> Practical Machine Learning by John Hopkins (Coursera)

Computer Science Fu

› CS50x on edX

Grasp end-to-end development

The things you build will be integrated into other systems

sas R python

Choose a first langu

› Open Source: R, Pyt

› Commercial: SAS, S

Learn Interactively

› R: DataCamp, tryR

› 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.

PostgreSQL

Apache CouchDB relax

4 Master data munging visualization and repo

□ Data cleaning and mungi

WHAT

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

TOOL

> Getting and by John Ho

DataWran

R data. dplyr

cassandra

MySQL

mongoDB

Learn more on databases via:

SQLZOO

MongoDB UNIVERSITY

datamonkey.pro

□ Data visualization

WHAT

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

TOOL

ggvis

D3

□ Reporting

WHAT

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

TOOL

+ a b

⦾S

R Ma

5 Level up with Big Data

When you start operating with data at the scale of the web, the fundamental approach
analysis must change. Most data scientists are working on problems that can't be run
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.

hadoop

MapReduce

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

Apache Spark i
speedy Swiss A
is a fast-runnin
analysis system
provides real-ti
processing func
Hadoop.

Spa

6 Get experience, practice and meet fellow data scientists

Practice makes perfect ...

kaggle
Meetup

Join in competitions

Meet fellow data scientists

Have a pet project

Dev in

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

8 Follow and engage wi the community

Sites to follow

- > DataTau
- > Kdnuggets
- > fivethirtyeight
- > datascience101
- > r-bloggers

Peop

- > Hi
- > Da
- > Na
- > dj

Need Data?

DataCamp
Learn data analysis for free, interactively
@DataCamp_com