THE UNIVERSITY OF CHICAGO

Kenneth C. Griffin Department of Economics
Office of Graduate Student Affairs
1126 East 59th Street, SHFE 511
Chicago, IL 60637
Telephone: 773.834.1972
Fax: 773.702.8490

Guide to Becoming a Data Scientist

The goal: be a moderately capable data scientist but an adeptly conversant one, with far greater business acumen leveraging your economics experience.

1) Read *Elements of Statistical Learning* - read chapters 1-10, 14-15 at least, and understand how the methods are similar (e.g., objective function minimization) / different (e.g., no closed form, no clustering of standard errors) from what you learned in econometrics (http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf). If limited time, read the baby-version instead http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf. I just did the baby version :)

2) Take Andrew Ng's Coursera class on Machine Learning - https://www.coursera.org/learn/machine-learning Attempt the assignments that seem relevant, but use R or Python [scientific libraries]. I did them all in R; prompts and solutions are here: https://github.com/faridcher/machine-learning-course.

3) Really learn R or Python [scientific libraries] - and use them in a literate research way (RMarkdown or Jupyter Notebook documents with comments around cacheable code). I learned R and RMarkdown.

4) Do some practice take-home challenges: https://datascientistjobinterview.com/

5) Use git for version control of code https://git-scm.com/

6) Learn basic Shell Scripting (aka Unix or BASH) commands for streamlining flow of analyses (data storage).

7) Learn basic data warehouse access commands (Hive / Pig from Hadoop; or Spark), or at least basic SQL. Incorporate the SQL / Hive code into R or Python.

8) Learn some about data infrastructure. I started making my way through this: https://blog.treasuredata.com/blog/2016/03/15/self-study-list-for-data-engineers-and-aspiring-data-architects/

9) BONUS: Complete your own research with:
i) Data accessed by API
ii) Scheduled data pull using a chron job or similar
iii) Use git to track your code revisions
iv) Have a shell script *.sh file that executes your entire work flow
v) A RMarkdown / Jupyter document with code + text

Kindly offered by Alison Rauh (Phd'11)