Data Science 101 - Survival Kit

Dr. Roy Sasson
sassonr@post.tau.ac.il
sasson.roy@idc.ac.il
roy.sasson@gmail.com

DATA ANALYSIS

SQL, Data Cleaning, Tableau, Reports, Data Visualization, R, Business Intelligence

---

MACHINE LEARNING

Supervised Learning
Regression, Deep Learning, SVM
Collaborative Filtering*

Unsupervised Learning
Pattern recognition,Clustering,
K-means, GMM, LDA, PCA

ARTIFICIAL INTELLIGENCE

Reinforcement Learning, Deep Learning,
Motion Planning, Robotics

Data Science 101 - Survival Kit

About this guide:
Many people want to learn how to work with data, either as part of their day job, or when considering a new career path. Luckily today the problem is not learning resources, rather than finding the *right* resources and the willpower to spend an evening or two per week and learning. All the resources in this guide were examined first-hand, and picked from many alternatives that are available online. But I'm sure there are others.
Focus in this guide is on videos and very short text-tutorials, as they seem to be the most time-efficient. *Beware* - Completing this guide will make you a data scientist only in theory. Nothing beats experience. In addition - most courses in this guide do not place an emphasis on mathematical or statistical theory, so that anyone can try them. This is a double-edged sword. On the one hand - it will get you onboarded quickly to the data science world, covering even advanced methods. On the other hand - there is some (high) glass ceiling in this profession which you cannot pass eventually without hard core math.
To help you prioritize between learning-investment and getting-started with your day-job, I've used the following color conventions:

***** Should-know before you pull data for your first data-driven project.

****** Learn from these resources a few hours a week after you already started your first project/s.

******* More advanced stuff (relevant for the more sophisticated data mining techniques).

******** learn in your spare time for Career Development, or in the case you want to work more efficiently.

If you are considering whether to learn R or Python - that mainly depends on your working environment. Familiarity with Python would help you along the way when you would want to raise your own data-driven services, but its harder. This guide focuses on R, which is easier to learn for non-engineers.

All resources here are either free, or require a symbolic charge (and worth it). Comments and suggestions are welcome roy.sasson@gmail.com

- What you need to know in order to become a proficient data scientist/analyst:

a. SQL for pulling data from databases, and from Hadoop or BigQuery
    i. Google the book “SQL for dummies”* and practice it.
    ii. Analytical functions** - great tutorial with examples.

b. Hive Syntax - sometimes a bit different than SQL. mostly similar
    i. Get a script from one of the other people in your office and use it as anchor*.
    ii. Linux command line syntax - spend a few hours learning** how to move around files and manipulate text using the command line.

c. Basic R syntax and data manipulation for descriptive statistics - mostly on datacamp. Worth the $50 (check with your manager about reimbursement) at least in this part of your training - make sure you complete the online exercises (ie - write code when requested. don't just watch the videos).

    DataCamp

    i. introduction to R*
    ii. intermediate R*
    iii. data analysis with R - the data.table way**
    iv. data manipulation in R with dplyr**
    v. cleaning data in R**
    vi. Data Table Cheat Sheet** - it will change your life

d. Data analysis with R**
    i. Meet with one of the other data scientists in your office**, and go over one of their R scripts. It will get you up to speed the fastest way.
    ii. Basic statistics** in R and AB testing courses are available online.

e. Basic statistics* - on coursera
   take courses 1,2 and 4 from Duke's Specialization

f. Data-unit-testing and documentation conventions*
   i. Always compare the results you get on your computer with some commonly-used reports in your company, for sanity checks.
   ii. Run a unit-test for each of your commands, no matter how simple it is. usually it would just be typing something like *summary(yourData$yourVariable)* and that's it...
   iii. Presenting data to stakeholders: here is a lecture I recently gave at TAU about Data Analysis best practices (in Hebrew).

g. Data Visualization with R*
   i. Learn ggplot. Nothing else. From there - take it with stack overflow.
   ii. Plotly R library*** - in case you really want to learn another plotting library.

h. Geo-based*** data analysis:
   i. Google S2 library

i. Machine learning - supervised learning:
   i. Linear regression** - the simple-to-use-and-understand workhorse of ML.
      coursera course - regression - university of Washington - wk 1-5
   ii. Logistic regression and ROC curve*** - when you want to predict 0 or 1 and are not satisfied with linear regression (which in many cases should be just fine).
      coursera course - classification - university of Washington
   iii. Classification trees*** - when your dependent and explanatory variables are mostly 0/1 or categorical + working with Logistic Regression becomes too complex

1. The basic CRAN (not the optimal method, but it is necessary to learn it for the general idea of classification trees)
2. Random Forest and Adaboost - usually the top performers in classification tasks
3. Learn the above from

   Coursera course - classification - university of Washington

j. Machine learning - unsupervised learning: When you are trying to understand patterns in your data, instead of making predictions. For example - what products people shop for in the same basket, movies that similar people watch, etc.)
   i. For all the below - take on Coursera

      Machine Learning - Clustering and Retrieval - wk1-5***
   ii. Similarity*** metrics (cosine, correlation, etc).
   iii. Clustering*** - K-means algorithm - the workhorse of unsupervised learning.
   iv. Clustering - more advanced techniques**** - necessary when the data is more complex (for example - an article is not only about medicine, but also about nutrition).
      1. GMM
      2. LDA
      3. Clustering - evaluation metrics - Urbana Champaign course in coursera has a good chapter on that

k. Deep learning**** (Considered state-of-the-art for Speech Recognition like Siri and Image Recognition) - take on Coursera:

   Andrew Ng's specialization on Coursera. can be audited for free.

l. Recommendation systems**** - How Facebook determines what to place on your newsfeed? How Netflix Recommends you with new movies?

i. Take this specialization on Coursera, University of Minnesota

m. Causal inference**** - Machine Learning only looks at correlations. It doesn't tell you what causes what ("the best doctors have the largest number of deceased patients" is an example for correlation != causality).
It is highly relevant in scenarios where you cannot have random assignment in AB tests. Econometrics courses are usually the best resource for understanding the difference between correlation and causality, as they usually rely on linear regression fundamentals, which you are likely to know.
My course is open for free (in Hebrew) on YouTube. There are many others online.

n. Introduction to AI (Artificial Intelligence)**** - take reinforcement learning class (Georgia University) on

o. Time series analysis and anomaly detection (to be added).

- Basic data analysis with Spark** - state-of-the-art in terms of heavy-lifting of data and data-analysis. requires a bit more technical savviness.
    a. Spark syntax guide.
    b. There is a Course on edX as well (did not take it).

- Keeping up-to-date with industry trends:****

    a. **A healthy research approach that could save you much time** - always assume that if you've thought of an idea, then it was already implemented and documented somewhere by the research community.
    Important Conferences - a great resource for getting ideas and implementations to research problems, most of them keep protocols online.

        i. KDD - focuses on ML and Data mining. The best conference in terms of practical combination of industry and machine learning
        ii. WWW - similar to KDD, but more hard-core ML

iii. WSDM - web search and data mining
iv. STRATA - O'reilly's conference arm. More emphasis on data engineering.
v. RecSys - recommendation systems (semi-academic, very innovative, but hard core math is usually applied in papers)

b. Recommended people/magazines to follow on Twitter (and yes.. you should have a Twitter account just for reading the cool stuff these folks share online! it's like a personal magazine editor..)

i. People:

- Hadley Wickham (R guru)
- Elon Musk
- Andrew Ng (founder of coursera, deep-learning guru)
- David Smith (R and data guru at MS)
- Dj patil (U.S. chief data scientist at office of science)
- John Myles White (Data Science at FB)
- Ben Lorica (Chief Data Scientist at O'Reilly)
- Jeff Weiner (CEO LinkedIn)
- Rasmus Baath (R guru)

ii. Magazines:

GIZMODO

WIRED

Geektime

Google Research

MIT Technology Review

Medium

(they have a section with great blog posts about Data Science)

c. Entertainment: watch HBO's Silicon Valley****

I like to say that it's more of a documentary rather than a comedy. Will teach you quite a lot about the startup world.