

7 Steps for Learning Data Mining and Data Science

[Share on facebook](#) [Share on linkedin](#)

91

[Previous post](#)

[Next post](#)

How to learn data mining and data science? I outline seven steps and point you to resources for becoming a data scientist.

By Gregory Piatetsky, Oct 10, 2013. [comments](#)

I am frequently asked - how to learn Data Mining and Data Science?

Here is my summary. Let me know what I missed and add your comments below.

You can best learn data mining and data science by doing, so start analyzing data as soon as you can!

However, don't forget to learn the theory, since you need a good statistical and machine learning foundation to understand what you are doing and to find real nuggets of value in the noise of Big Data.

Here are 7 steps for learning data mining and data science. Although they are numbered, you can do them in parallel or in a different order.

1. Languages: Learn R, Python, and SQL
2. Tools: Learn how to use data mining and visualization tools
3. Textbooks: Read introductory textbooks to understand the fundamentals
4. Education: watch webinars, take courses, and consider a certificate or a degree in data science
5. Data: Check available data resources and find something there
6. Competitions: Participate in data mining competitions
7. Interact with other data scientists, via social networks, groups, and meetings

Also, don't forget to subscribe to [KDnuggets News](#) bi-weekly email and follow [@kdnuggets](#) - voted [Top Big Data Twitter](#) - for latest news on Analytics, Big Data, Data Mining, and Data Science.

Here I use Data Mining and Data Science interchangeably - see my presentation [Analytics Industry Overview](#), where I look at evolution and popularity of different terms like Statistics, Knowledge Discovery, Data Mining, Predictive Analytics, Data Science, and Big Data.

1. Learning Languages

Recent KDnuggets Poll found that [the most popular languages for data mining](#) are R, Python, and SQL. There are many resources for each, for example

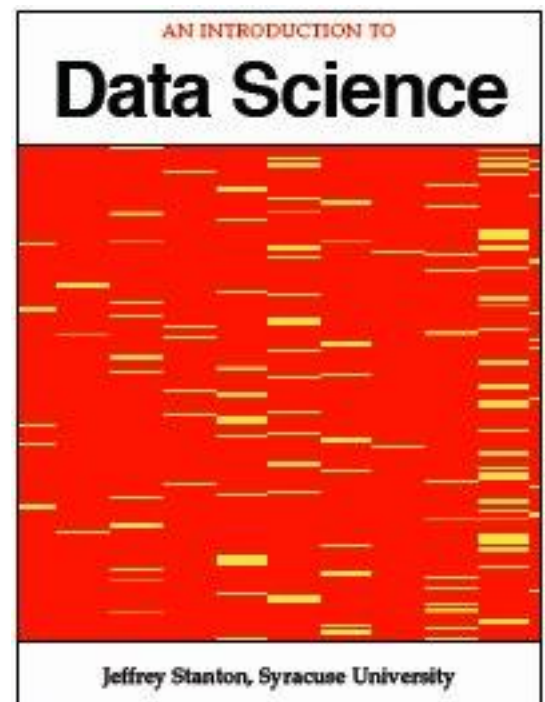
- [Free e-book on Data Science with R](#)
- [Getting Started With Python For Data Science](#)
- [Python for Data Analysis: Agile Tools for Real World Data](#)
- [An indispensable Python : Data sourcing to Data science.](#)
- [W3 Schools Learning SQL](#)

2. Tools: Data Mining, Data Science, and Visualization Software

There are many [data mining tools](#) for different tasks, but it is best to learn using a data mining suite which supports the entire process of data analysis.

You can start with open source (free) tools such as [KNIME](#), [RapidMiner](#), and [Weka](#).

However, for many analytics jobs you need to know [SAS](#), which is the leading commercial tool and widely used.



Other popular [Analytics and Data Mining Software](#) include MATLAB, StatSoft STATISTICA, Microsoft SQL Server, Tableau, IBM SPSS Modeler, and Rattle.

Visualization is an essential part of any data analysis - learn how to use Microsoft Excel (good for many simpler tasks), [R graphics](#), (especially [ggplot2](#)), and also [Tableau](#) - an excellent package for visualization. Other good visualization tools include TIBCO Spotfire and Miner3D.

3. Textbooks

There are many [data mining and data science textbooks](#) available, but you can check these

- [Data Mining and Analysis: Fundamental Concepts and Algorithms, free PDF download \(draft\)](#), by Mohammed Zaki and Wagner Meira Jr.
- [Data Mining: Practical Machine Learning Tools and Techniques](#), by Ian Witten, Eibe Frank, and Mark Hall, from the authors of Weka, and using Weka extensively in examples.
- [The Elements of Statistical Learning](#), Data Mining, Inference, and Prediction, by Trevor Hastie, Robert Tibshirani, Jerome Friedman - great introduction for mathematically oriented
- [LIONbook: Learning and Intelligent Optimization](#), by Roberto Battiti and Mauro Brunato, freely available on the web, chapter by chapter.
- [Mining of Massive Datasets Book](#), by A. Rajaraman, J. Ullman.
- [StatSoft Electronic Statistics Textbook™ \(free\)](#), includes many data mining topics

4. Education: Webinars, Courses, Certificates, and Degrees

You can start by watching some of the many free [webinars and webcasts](#) on latest topics in Analytics, Big Data, Data Mining, and Data Science.

There are also many online courses, short and long, many of them free - see [KDnuggets online education directory](#).

Check in particular these courses:

- [Machine Learning](#), at Coursera, taught by Andrew Ng
- [Learning from Data](#) at edX, taught by Caltech professor Yaser Abu-Mostafa,
- [Open Online Course in Applied Data Science](#), from Syracuse iSchool
- [Data Mining with Weka](#), free online course.
- check also free online slides from my [Data Mining Course](#), a semester-long introductory course in Data Mining.

Finally, consider getting [Certificates in Data Mining, and Data Science](#) or advanced degrees, such as MS in Data Science - see KDnuggets directory for [Education in Analytics, Data Mining, and Data Science](#).

5. Data

You will need data to analyze - see KDnuggets directory of [Datasets for Data Mining](#), including

- [Government, Federal, State, City, Local and public data sites and portals](#)
- [Data APIs, Hubs, Marketplaces, Platforms, Portals, and Search Engines](#).
- [Free Public Datasets](#)

6. Competitions

Again, you will best learn by doing, so participate in [Kaggle competitions](#) - start with beginner competitions, such as [Predicting Titanic Survival using Machine Learning](#)

7. Interact: Meetings, Groups, and Social Networks

You can join many peer groups - see [Top 30 LinkedIn Groups for Analytics, Big Data, Data Mining, and Data Science](#).

[AnalyticBridge](#) is an active community for Analytics and Data Science.

You can attend some of the many [Meetings and Conferences on Analytics, Big Data, Data Mining, Data Science, & Knowledge Discovery](#).

Also, consider joining [ACM SIGKDD](#), which organizes the annual KDD conference - the leading research conference in the field.

More ...