If you're a budding data scientist, a manager who wants to learn, or someone who wants to get started with data science or engineering, there is a chance you might become overwhelmed with all the things you'll need to learn! In fact, due to the scarcity of actual learning materials, one really needs to cobble together a do-it-yourself curriculum in data science.

Here is a curated list of data science and data engineering links and publications to help get you started. This list is by no means comprehensive; it's just a starting point.

Books, online and downloadable guides:

Background:

Milton and Arnold. **Introduction to Probability and Statistics.** McGraw Hill, 1995. ISBN 0-07-113535-9

Peter Kennedy. **A Guide to Econometrics.** MIT Press, 2003. ISBN: 978-0-262-61183-1

The Data Science Handbook. http://www.thedatasciencehandbook.com/#get-the-book

The Data Analytics Handbook. https://www.teamleada.com/handbook

Programming:

Wes McKinney. **Python for Data Analysis.** O'Reilly, 2013. ISBN: 978-1-449-31979-3

Michael Crawley. Statistics. **An Introduction using R.** Wiley, 2015. ISBN: 978-1-118-94109-6

Capriolo, Wampler and Rutherglen. **Programming Hive, First Edition.** O'Reilly, 2012. ISBN: 978-1-449-31933-5

Treasure Data's HIVE guide:
http://get.treasuredata.com/hive-guide

Introduction to Spark with Python:
http://www.kdnuggets.com/2015/11/introduction-spark-python.html

Learn Python the Hard Way:
http://learnpythonthehardway.org/book/

Infrastructure:

Tom White. **Hadoop, The Definitive Guide.** O'Reilly, 2012. ISBN: 978-1-449-31152-0
Eric Sammer. **Hadoop Operations.** O'Reilly, 2012. ISBN: 978-1-449-32705-7
Karau, et al. **Learning Spark: Lightning-Fast Big Data Analysis.** First Edition. O'Reilly, 2015. ISBN: 978-1-449-35862-4

Treasure Data's Amazon Redshift COPY command guide:
http://get.treasuredata.com/Redshift_COPY_Command_Guide-Request.html

Links:

24 Data Science Resources to Keep Your Finger on the Pulse:
http://blog.udacity.com/2014/12/24-data-science-resources-keep-finger-pulse.html

Treasure Data's Introduction to Data Science:
http://blog.treasuredata.com/blog/2015/06/23/data-science-101-interactive-analysis-with-jupyter-pandas-and-treasure-data/

Treasure Data's Introduction to Data Ingestion from (Python) Apps:
http://blog.treasuredata.com/blog/2015/04/24/python-for-aspiring-data-nerds/

Treasure Data's Blog: Learn SQL by Calculating Customer Lifetime Value:
http://blog.treasuredata.com/blog/2014/12/05/learn-sql-by-calculating-customer-lifetime-value-part-1/

Algorithmia's Blog on Consuming Algorithms via REST APIs in Python:
http://blog.algorithmia.com/post/115255527924/accessing-algorithmia-through-python

Plus...

Analytics

- Mixpanel Blog
- Amplitude Blog
- KISSMetrics Blog
- RJMetrics Blog
- Mode Analytics Blog
- Periscope Blog
- Google Analytics Blog
- GameAnalytics Blog

Data Management

- Curt Monash's blog
- O'Reilly Radar Data
- Ad Age Data (adtech)

Databases

- Aphyr's Call Me Maybe

Infrastructure as a Service

- AWS Blog
- AWS Big Data Blog

Platform as a Service

- Heroku Blog

Software as a Service

- SaaStr (more to understand SaaS business models)
- David Skok's blog (again, more to understand how SaaS business work)
- Marketo Blog (trying to integrate with Marketo)
- Salesforce Developer Blog (we integrate with them)

Workflow Management

- Airflow
- Luigi

Data Quality Management

- Trifacta Blog
- Paxata Blog

Data Collection/Ingestion/Stream Processing

- Fluentd
- Embulk
- Elastic (maker of Elasticsearch, Logstash and Kibana)
- Confluent Blog (maker of Apache Kafka)
- Databricks Blog (maker of Apache Spark)

Machine Learning

- ŷhat blog
- Domino Data Lab
- Data Tau (aggregator)

Treasure Data's Data Science Reading List

+1.866.899.5386
info@treasuredata.com
2565 Leghorn St.
Mountain View, CA 94043