# RoadMap Of Data Science

*Lesson Plan*

**Author**: Hayder Zaeem

**Date**: 2020

**GitHub**: hayderzaeim

**Twitter**: hayderzaeem

**Instagram**: hayder_zaeem

**LinkedIn**: Hayder Zaeem

© 2020 Hayder Zaeem. All rights reserved. 

## 1. Mathematics and Statistics

**Definition:** Mathematics and statistics form the foundational pillars of data science, providing the theoretical framework for analyzing and interpreting data.

- **Linear Algebra:** Essential for operations on vectors and matrices, crucial in machine learning algorithms like PCA and SVD.
- **Calculus:** Provides tools for optimization and understanding rate of change, applied in gradient descent for model training.
- **Probability:** Fundamental for modeling uncertainty and making predictions, foundational to Bayesian inference and probabilistic models.
- **Statistics:** Techniques for summarizing data, making inferences, and testing hypotheses, pivotal in experimental design and model evaluation.

## 2. Programming and Software Engineering

**Definition:** Proficiency in programming languages and software engineering practices enables effective implementation and deployment of data science solutions.

- **Programming Languages:** Python and R are widely used for data manipulation, analysis, and modeling. SQL for querying databases.
- **Software Development:** Includes version control (Git), testing, and debugging for maintaining code quality and collaboration.
- **Data Structures and Algorithms:** Knowledge of efficient data storage and retrieval methods, crucial for optimizing data processing pipelines.

### 3. Data Management and Data Wrangling

**Definition:** Involves acquiring, cleaning, and organizing data to prepare it for analysis and modeling.

- **Data Collection:** Methods such as web scraping and APIs to gather data from various sources.
- **Data Cleaning:** Handling missing values, outliers, and inconsistencies to ensure data quality.
- **Database Management:** Using relational (SQL) and NoSQL databases for storing and retrieving structured and unstructured data.

### 4. Exploratory Data Analysis (EDA)

**Definition:** EDA involves visualizing and summarizing data to uncover patterns, anomalies, and relationships that guide further analysis.

- **Data Visualization:** Tools like Matplotlib and ggplot2 for creating charts, graphs, and plots to explore data distributions and trends.
- **Summary Statistics:** Calculation of mean, median, standard deviation, etc., to summarize data characteristics.
- **Pattern Recognition:** Identification of recurring structures or features in data using statistical methods and visualization techniques.

### 5. Machine Learning

**Definition:** Machine learning focuses on developing algorithms that learn patterns and make predictions from data.

- **Supervised Learning:** Training models using labeled data for tasks like regression and classification.
- **Unsupervised Learning:** Discovering patterns and structures in unlabeled data through clustering and dimensionality reduction.

- **Reinforcement Learning:** Training models to make sequences of decisions through interaction with an environment.
- **Model Evaluation and Optimization:** Techniques such as cross-validation and hyperparameter tuning to improve model performance.

### 6. Deep Learning

**Definition:** Subset of machine learning focusing on neural networks with multiple layers, capable of learning complex representations from large datasets.

- **Neural Networks:** Mimic the human brain's structure, used in applications like image and speech recognition.
- **CNNs and RNNs:** Specialized architectures for tasks involving spatial and sequential data, respectively.
- **GANs:** Frameworks for generating new data instances that resemble the training data distribution.
- **Deep Reinforcement Learning:** Integrating deep learning with reinforcement learning for complex decision-making tasks.

### 7. Natural Language Processing (NLP)

**Definition:** NLP involves processing and analyzing human language data to extract meaning and insights.

- **Text Preprocessing:** Tokenization, stemming, and lemmatization to prepare text data for analysis.
- **Sentiment Analysis:** Determining the sentiment or opinion expressed in text data.
- **Named Entity Recognition (NER):** Identifying and categorizing named entities such as names, dates, and locations in text.

- **Language Models:** Advanced models like BERT and GPT for tasks such as text generation and language understanding.

### 8. Big Data Technologies

**Definition:** Encompasses tools and frameworks for processing and analyzing large volumes of data efficiently.

- **Distributed Computing:** Platforms like Hadoop and Spark for processing massive datasets across clusters of computers.
- **Big Data Storage:** Techniques like HDFS for distributed storage of data across multiple nodes.
- **NoSQL Databases:** Scalable databases like Cassandra and HBase for handling unstructured and semi-structured data.

### 9. Cloud Computing

**Definition:** Utilizing cloud platforms and services to store, manage, and analyze data over the internet.

- **Cloud Platforms:** AWS, Google Cloud, and Azure provide infrastructure and services for deploying and scaling data science applications.
- **Cloud Services:** Offerings include data storage, compute resources, and machine learning services for scalable and cost-effective data processing.

### 10. Data Engineering

**Definition:** Focuses on designing and building pipelines to extract, transform, and load data for analysis and modeling.

- **ETL Processes:** Extracting data from various sources, transforming it to fit operational needs, and loading it into data warehouses or databases.
- **Data Pipeline Design:** Architecting workflows to automate data processing tasks and ensure data integrity and reliability.
- **Real-Time Data Processing:** Tools like Kafka and Flink for handling continuous streams of data and processing them in real-time.

### 11. Domain Knowledge and Ethics

**Definition:** Understanding specific industries or domains where data science solutions are applied, along with ethical considerations in data use.

- **Domain-Specific Applications:** Applying data science techniques to fields such as finance, healthcare, marketing, etc., to solve domain-specific challenges.
- **Data Ethics:** Addressing issues of privacy, bias, and fairness in data collection, analysis, and decision-making algorithms.

### 12. Advanced Topics and Research

**Definition:** Involves cutting-edge research and advanced techniques pushing the boundaries of data science.

- **Advanced Machine Learning:** Meta-learning, few-shot learning, and advanced optimization techniques.
- **Advanced Deep Learning:** Transformer architectures, attention mechanisms, and generative models.
- **Interdisciplinary Applications:** Applying data science to emerging fields like bioinformatics and computational social science.
- **Cutting-Edge Research:** Exploring areas such as AI safety, quantum computing applications in data science, and beyond.

## 13. Capstone Projects and Practical Experience

**Definition:** Hands-on projects and real-world applications to solidify skills and demonstrate proficiency in data science.

- **Industry Projects:** Collaborating with organizations to solve real-world problems using data science techniques.
- **Academic Research Projects:** Contributing to academic knowledge through research in data science methodologies and applications.
- **Competitions:** Platforms like Kaggle for participating in data science competitions to showcase skills and learn from peers.

## 14. Professional Development

**Definition:** Continuous learning, networking, and career advancement in the field of data science.

- **Networking and Community Involvement:** Attending conferences, meetups, and online forums to connect with professionals and stay updated on industry trends.
- **Publishing Research:** Contributing findings to journals and conferences to advance knowledge and establish expertise in specific areas of data science.
- **Certifications and Continuous Learning:** Pursuing certifications, online courses, and workshops to stay current with evolving tools, techniques, and best practices in data science.

## 1. Mathematics and Statistics

- **Tools:**
  - **R:** Widely used for statistical analysis and visualization.
  - **Python Libraries:**
    - **NumPy:** For numerical computations and linear algebra operations.
    - **SciPy:** For advanced mathematical functions and statistics.
    - **StatsModels:** For statistical modeling and hypothesis testing.
  - **Mathematica:** For symbolic mathematics and numerical computation.
  - **MATLAB:** For numerical computing and algorithm development.

## 2. Programming and Software Engineering

- **Tools:**
  - **Python:** General-purpose programming language with extensive data science libraries.
  - **R:** Statistical programming language.
  - **SQL:** For managing and querying relational databases.
  - **Git:** Version control system for tracking changes in code.
  - **Jupyter Notebook:** Interactive coding environment for Python.
  - **PyCharm:** Integrated Development Environment (IDE) for Python.
  - **RStudio:** IDE for R.

- **VS Code:** Versatile code editor with extensions for various languages.

### 3. Data Management and Data Wrangling

- **Tools:**
  - **Pandas (Python):** For data manipulation and analysis.
  - **Dplyr (R):** For data manipulation.
  - **SQL:** For querying relational databases.
  - **Apache Hadoop:** Framework for distributed storage and processing of large data sets.
  - **Apache Spark:** Unified analytics engine for big data processing.
  - **Talend:** Data integration tool.
  - **Alteryx:** Data preparation and blending tool.
  - **Excel:** For simple data manipulation and analysis.

### 4. Exploratory Data Analysis (EDA)

- **Tools:**
  - **Matplotlib (Python):** For plotting and visualization.
  - **Seaborn (Python):** For statistical data visualization.
  - **Plotly (Python/R):** For interactive plots.
  - **ggplot2 (R):** For creating complex plots from data in a data frame.
  - **Tableau:** For creating interactive visualizations.
  - **Power BI:** Business analytics service with visualization tools.

### 5. Machine Learning

- **Tools:**
  - **Scikit-learn (Python):** For machine learning algorithms and data mining.

- **TensorFlow (Python):** Open-source library for machine learning and deep learning.
- **Keras (Python):** High-level neural networks API.
- **XGBoost (Python/R):** For gradient boosting.
- **LightGBM (Python/R):** Gradient boosting framework.
- **Caret (R):** For machine learning.
- **H2O.ai:** Open-source platform for AI and machine learning.

### 6. Deep Learning

- **Tools:**
  - **TensorFlow (Python):** Open-source platform for machine learning.
  - **Keras (Python):** API for building and training deep learning models.
  - **PyTorch (Python):** Deep learning framework.
  - **Caffe:** Deep learning framework made with expression, speed, and modularity in mind.
  - **MXNet:** Deep learning framework designed for efficiency and flexibility.
  - **Theano:** Library for defining, optimizing, and evaluating mathematical expressions.

### 7. Natural Language Processing (NLP)

- **Tools:**
  - **NLTK (Python):** Toolkit for working with human language data.
  - **SpaCy (Python):** Industrial-strength NLP library.
  - **Gensim (Python):** For topic modeling and document similarity.
  - **BERT (Python):** Pre-trained NLP model.
  - **GPT (Python):** Generative Pre-trained Transformer model.

- **Stanford NLP:** Suite of NLP tools provided by Stanford University.

## 8. Big Data Technologies

- **Tools:**
  - **Apache Hadoop:** Framework for distributed storage and processing.
  - **Apache Spark:** Unified analytics engine for big data processing.
  - **Kafka:** Distributed streaming platform.
  - **Flink:** Stream processing framework.
  - **HDFS:** Hadoop Distributed File System for storage.
  - **Hive:** Data warehouse software for reading, writing, and managing large datasets.
  - **Cassandra:** NoSQL database.
  - **HBase:** Non-relational distributed database modeled after Google's Bigtable.

## 9. Cloud Computing

- **Tools:**
  - **AWS (Amazon Web Services):** Comprehensive cloud computing platform.
  - **Google Cloud Platform:** Suite of cloud computing services.
  - **Microsoft Azure:** Cloud computing service.
  - **Databricks:** Unified data analytics platform.
  - **Snowflake:** Cloud data platform.
  - **Kubernetes:** For container orchestration.
  - **Docker:** For containerization.
  - **Terraform:** For infrastructure as code.

## 10. Data Engineering

- **Tools:**
    - **Apache Airflow:** Platform for programmatically authoring, scheduling, and monitoring workflows.
    - **Kafka:** Distributed streaming platform.
    - **NiFi:** Data integration tool.
    - **Apache Beam:** Unified programming model for batch and streaming data processing.
    - **AWS Glue:** Fully managed ETL service.
    - **dbt (Data Build Tool):** For data transformation.
    - **Snowflake:** Data warehousing solution.

## 11. Domain Knowledge and Ethics

- **Tools:**
    - **Industry-specific software:** Depending on the domain (e.g., SAS for healthcare, Bloomberg Terminal for finance).
    - **Ethics training platforms:** Various online ethics training courses.

## 12. Advanced Topics and Research

- **Tools:**
    - **Papers with Code:** Repository of research papers and their implementations.
    - **ArXiv:** Repository of research papers.
    - **TensorFlow Research Cloud:** Program providing access to cloud TPUs for research.
    - **Google Colab:** For executing Python in the cloud, ideal for research prototypes.

## 13. Capstone Projects and Practical Experience

- **Tools:**
  - **Kaggle:** Platform for data science competitions.
  - **GitHub:** For managing and sharing code.
  - **Google Colab:** For executing Python in the cloud, sharing notebooks.
  - **Jupyter Notebook:** For creating and sharing documents that contain live code, equations, visualizations.

## 14. Professional Development

- **Tools:**
  - **LinkedIn Learning:** Online learning platform with courses.
  - **Coursera:** Online courses, certifications.
  - **edX:** Online courses, certifications.
  - **Udacity:** Nanodegrees and career services.
  - **Meetup:** Platform for finding and building local communities.
  - **Kaggle:** For competitions and community involvement.
  - **Professional societies:** Such as IEEE, ACM, and other industry-specific organizations.

## 1. Mathematics and Statistics

**Learning Resources:**

- Khan Academy - Linear Algebra, Calculus, Statistics
- MIT OpenCourseWare - Mathematics for Computer Science

- Coursera - Mathematics for Machine Learning Specialization

**Certificate Courses:**

- edX - Mathematics for Data Science Professional Certificate
- Coursera - Statistics with R Specialization

**University Programs:**

- Stanford University - MS in Statistics
- University of California, Berkeley - Master of Information and Data Science (MIDS)

## 2. Programming and Software Engineering

**Learning Resources:**

- Codecademy - Python, SQL, R Programming
- Udacity - Programming for Data Science Nanodegree
- GitHub Learning Lab - Git and GitHub Basics

**Certificate Courses:**

- Coursera - Python for Everybody Specialization
- edX - Software Development for Data Scientists

**University Programs:**

- Massachusetts Institute of Technology (MIT) - Master of Business Analytics
- University of Washington - Master of Science in Data Science

## 3. Data Management and Data Wrangling

**Learning Resources:**

- DataCamp - Data Manipulation with Python and R
- Udemy - SQL for Data Science
- LinkedIn Learning - Data Cleaning and Preprocessing Techniques

**Certificate Courses:**

- Coursera - Data Engineering with Google Cloud Professional Certificate
- edX - MongoDB for Developers

**University Programs:**

- Carnegie Mellon University - Master of Science in Information Technology - Business Intelligence and Data Analytics
- University of Illinois at Urbana-Champaign - Master of Computer Science in Data Science

## 4. Exploratory Data Analysis (EDA)

**Learning Resources:**

- Data Visualization with Python and Matplotlib - Book by Jake VanderPlas
- Udacity - Data Visualization Nanodegree
- Tableau - Online Training and Tutorials

**Certificate Courses:**

- Coursera - Data Visualization with Tableau Specialization
- edX - Data Science: Visualization

**University Programs:**

- University of Michigan - Master of Applied Data Science

- Columbia University - Master of Science in Data Science

## 5. Machine Learning

**Learning Resources:**

- Andrew Ng's Machine Learning Course on Coursera
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Book by Aurélien Géron
- Kaggle - Machine Learning and Data Science Competitions

**Certificate Courses:**

- Coursera - Machine Learning by Stanford University
- edX - Applied Machine Learning

**University Programs:**

- University of California, Los Angeles (UCLA) - Master of Science in Data Science
- University of Texas at Austin - Master of Science in Data Science and Analytics

## 6. Deep Learning

**Learning Resources:**

- Deep Learning Specialization by Andrew Ng on Coursera
- Dive into Deep Learning - Online Book by Aston Zhang et al.
- TensorFlow and PyTorch Documentation and Tutorials

**Certificate Courses:**

- Udacity - Deep Learning Nanodegree

- Coursera - Deep Learning Specialization by deeplearning.ai

**University Programs:**

- New York University (NYU) - Master of Science in Data Science - Deep Learning Specialization
- University of Toronto - Master of Science in Applied Computing - Artificial Intelligence and Deep Learning Track

## 7. Natural Language Processing (NLP)

**Learning Resources:**

- Natural Language Processing with Python - Book by Steven Bird, Ewan Klein, and Edward Loper
- Stanford NLP Group - Online Resources and Tutorials
- SpaCy - Documentation and Tutorials

**Certificate Courses:**

- Coursera - Natural Language Processing Specialization
- edX - Natural Language Processing with Python

**University Programs:**

- Johns Hopkins University - Master of Science in Data Science and NLP
- University of Washington - Master of Science in Computational Linguistics

## 8. Big Data Technologies

**Learning Resources:**

- Hadoop: The Definitive Guide - Book by Tom White

- Spark Documentation and Tutorials
- DataBricks Academy - Online Training for Apache Spark

**Certificate Courses:**

- Coursera - Big Data Specialization
- edX - Apache Kafka Series

**University Programs:**

- Georgia Institute of Technology - Master of Science in Analytics - Big Data Track
- University of California, San Diego - Master of Data Science with a Specialization in Big Data

### 9. Cloud Computing

**Learning Resources:**

- AWS Training and Certification
- Google Cloud Training
- Microsoft Learn - Azure

**Certificate Courses:**

- AWS Certified Solutions Architect - Associate
- Google Cloud Professional Cloud Architect Certification
- Microsoft Certified: Azure Solutions Architect Expert

**University Programs:**

- University of Maryland Global Campus - Master's in Cloud Computing Architecture

- Northwestern University - Master of Science in Information Technology with Cloud Computing Specialization

## 10. Data Engineering

**Learning Resources:**

- Data Engineering Cookbook - Online Book by Andreas Kretz
- Apache Airflow Documentation
- Kafka Tutorials and Documentation

**Certificate Courses:**

- Coursera - Data Engineering on Google Cloud Professional Certificate
- edX - Data Engineering with Python

**University Programs:**

- University of Southern California - Master of Science in Computer Science (Data Engineering Track)
- University of Chicago - Master of Science in Analytics - Data Engineering Track

## 11. Domain Knowledge and Ethics

**Learning Resources:**

- Data Science for Business - Book by Foster Provost and Tom Fawcett
- Online Ethics Courses - Coursera, edX
- Industry-specific webinars and seminars

**Certificate Courses:**

- Coursera - Business Analytics Specialization

- edX - Data Science and Ethics

**University Programs:**

- Harvard University - Master of Science in Data Science - Business Analytics Track
- University of Pennsylvania - Master of Science in Data Science - Social Science Analytics Track

## 12. Advanced Topics and Research

**Learning Resources:**

- Papers with Code - Platform for AI research papers and implementations
- ArXiv - Repository of research papers in various fields including AI and data science
- Online conferences and webinars by AI research organizations

**Certificate Courses:**

- Coursera - Advanced Machine Learning Specialization
- edX - Quantum Machine Learning

**University Programs:**

- Stanford University - PhD in Computer Science with a focus on AI and Machine Learning
- Massachusetts Institute of Technology (MIT) - PhD in Electrical Engineering and Computer Science - AI and Machine Learning

## 13. Capstone Projects and Practical Experience

**Learning Resources:**

- Kaggle - Data Science Competitions and Datasets
- Open Source Projects on GitHub
- Internships and Industry Projects

**Certificate Courses:**

- Coursera - Applied Data Science Capstone Project
- edX - Data Science Capstone Project

**University Programs:**

- Carnegie Mellon University - Master of Computational Data Science - Practicum in Data Science
- University of California, Irvine - Master of Science in Data Science - Practicum and Projects

## 14. Professional Development

**Learning Resources:**

- LinkedIn Learning - Career Development Courses
- Udacity - Career Services and Job Placement Support
- Data Science Conferences and Meetups

**Certificate Courses:**

- Coursera - Professional Certificate in Data Science
- edX - MicroMasters Program in Data Science

**University Programs:**

- Northwestern University - Master of Science in Analytics - Career Development Services

- Duke University - Master of Interdisciplinary Data Science - Professional Development Workshops