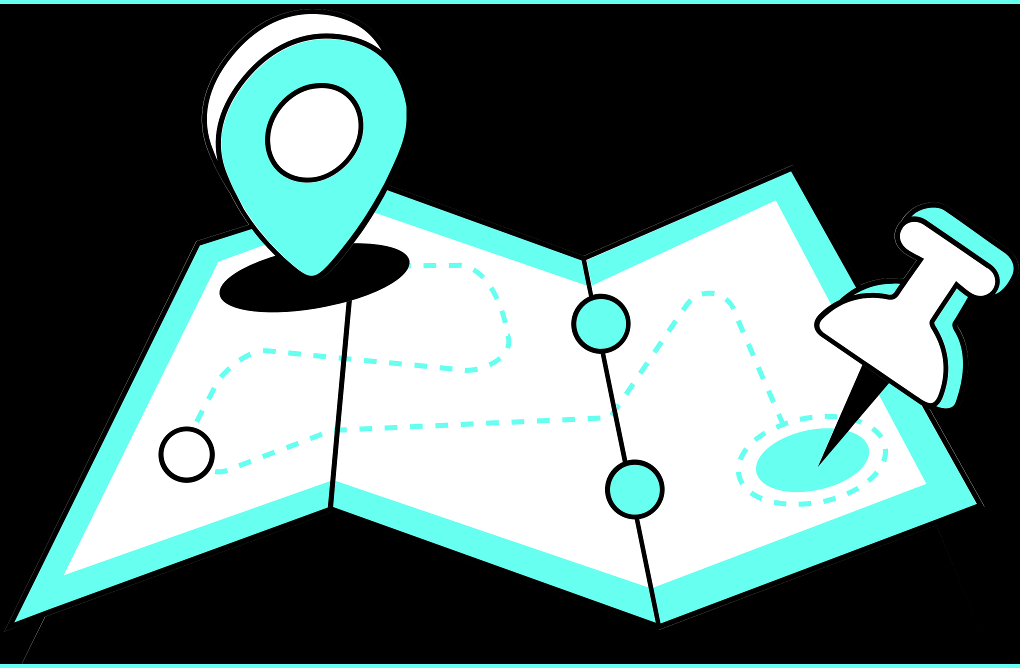
THE COMPLETE

DATA SCIENCE ROADMAP

2025



Go From Zero to a Data Scientist
in 12 Months

Juma Shafara

Hi! I am Juma Shafara, a data scientist at Raising The Village and Instructor at DATAIDEA with over 4 years of experience.

Over the past 2 years, I've had the privilege of teaching hundreds of people how to code and become professional data scientists through my startup, DATAIDEA and online courses.

It's my mission to make programming accessible to everyone. Join me on this journey and unlock your potential in the world of coding!

www.dataidea.org

Table of Content

Introduction

This guide is designed to help you navigate the essential skills needed to become a successful data scientist. Whether you're just starting out or looking to enhance your existing skills, this roadmap will provide a clear and structured path.

Target Audience

This guide is for:

•**Beginners** who want to know what they need to learn to land a job in data science

•**Experienced individuals** looking to level up their skills and fill in gaps in their knowledge

Resources

For detailed tutorials and full courses, check out the following resources:

YouTube: https://www.youtube.com/@dataidea-science
Full Courses: https://science.dataidea.org/

Roadmap Overview

Below is a comprehensive table listing all the essential skills needed to become a proficient data scientist, along with the estimated time required to learn each skill.

Keep in mind that the time needed to learn each skill can vary for everyone. These estimates are based on dedicating 3 to 5 hours of study every day.

Use this roadmap to guide your learning journey and track your progress as you build a strong foundation in data science.

Skill	Est. Time	Learning Phase
Programming (Python)	1 to 2 months	Beginner
Version Control (Git)	1 to 2 weeks	Beginner
Data Structures & Algorithms	1 to 2 months	Beginner
SQL	1 to 2 months	Beginner
Mathematics and Statistics	2 to 3 months	Beginner
Data Collection and Visualization	1 to 2 months	Intermediate
Machine Learning Fundamentals	2 to 3 months	Intermediate
Deep Learning	2 to 3 months	Advanced
Specialization (NLP or Computer Vision)	2 to 3 months	Advanced
Big Data (Optional)	2 to 3 months	Advanced
Total	**12 to 20 months**	

Python

Python is a highly popular language for data science, known for its simplicity, readability, and extensive library support. It's widely used for data analysis, visualization, and building machine learning models.

Estimated time: 2 months

Essential Concepts

- **Python Fundamentals**

  •Variables and data types

  •Loops (for, while) and conditional statements (if, elif, else)

  •Functions and scope

- **Data Structures**

  •Arrays, lists, tuples and sets

  •Stacks and queues

  •Dictionaries

  •Comprehensions

  •Generator expressions

  Exception Handling

- •Handling exceptions with try/except

  •Raising exceptions

  Functional Programming

  •Lambda functions
  •Map, reduce, filter

- **Object-oriented Programming**
    - Classes and objects
    - Inheritance and polymorphism
- **Modules and packages**
    - Creating modules
    - Managing packages with pip and pipenv
    - Virtual environments
- **Python Standard Library**:
    - Working with paths, files, and directories
    - Working with CSV and JSON files
    - Working with Date/time
    - Generating random values
- **Familiarity with data science libraries**
    - NumPy
    - Pandas
    - Matplotlib

Version Control (Git)

Git is a version control system that is crucial for managing code and collaboration in data science projects. It allows you to track changes, collaborate with others, and maintain the integrity of your codebase.

Estimated time: 1–2 weeks

Learning resources: Full Course

Essential Concepts

- **Setup and Configuration**: init, clone, config
- **Staging**: status, add, rm, mv, commit, reset
- **Inspect and Compare**: log, diff, show
- **Branching**: branch, checkout, merge
- **Remote Repositories**: remote, fetch, pull, push
- **Temporary Commits**: stash
- **GitHub**: fork, pull request, code review

Data Structures & Algorithms

Understanding data structures and algorithms is crucial for optimizing code and solving complex problems efficiently. This knowledge is fundamental for technical interviews and real-world data science tasks.

Estimated Time: 1–2 months

Learning resources: Full Course

Essential Concepts

Big O Notation Arrays and Linked Lists Stacks and Queues Hash Tables

Trees and Graphs

- Binary trees
- AVL trees
- Heaps
- Tries
- Graphs

Sorting Algorithms

- Bubble sort
- Selection sort
- Insertion sort
- Merge sort
- Quick sort
- Counting sort
- Bucket sort

- **Searching algorithms**
    - Linear search
    - Binary search
    - Ternary search
    - Jump search
    - Exponential search
- **String Manipulation Algorithms**
    - Reversing a string
    - Reversing words
    - Rotations
    - Removing duplicates
    - Most repeated character
    - Anagrams
    - Palindrome
- **Recursion**

SQL

SQL (Structured Query Language) is essential for querying and managing data in relational databases. It's a fundamental skill for any data scientist working with structured data.

Estimated time: 1-2 months

Learning resources: Full Course

Essential Concepts

- **Basic Operations**
    - Querying data (SELECT)
    - Modifying data (INSERT, UPDATE, DELETE)
    - Filtering data (WHERE, IN, BETWEEN, LIKE, IS NULL, REGEXP)
    - Logical operators (AND, OR, NOT)
    - Sorting and limiting data (ORDER BY, LIMIT)
- **Complex Queries**
    - Joins (INNER, OUTER, SELF, NATURAL, CROSS)
    - Aggregate functions (MAX, MIN, AVG, SUM, COUNT)
    - Grouping data (GROUP BY, HAVING, ROLLUP)
    - Subqueries
- **Views**
- **Stored Procedures and Functions**
- **Triggers and Events Transactions**

  - Transaction isolation levels
  - BEGIN, COMMIT, ROLLBACK

- **Database Design**
  - Normalization
  - Database integrity with primary keys, foreign keys, and constraints

- **Indexes**
- **Security and Permissions**: Managing users and privileges

Mathematics and Statistics

Mathematics and statistics are fundamental for understanding data science concepts. They provide the theoretical foundation for data analysis and machine learning algorithms.

Estimated Time: 2–3 months

Essential Concepts

- **Linear Algebra**
    - Vectors and matrices
    - Matrix operations
    - Eigenvalues and eigenvectors
    - Singular Value Decomposition (SVD)

- **Calculus**
    - Derivatives and gradients
    - Partial derivatives
    - Chain rule
    - Integrals

- **Probability**
    - Probability distributions
    - Bayes' theorem
    - Random variables
    - Expectation and variance

- **Statistics**
    - Descriptive statistics (mean, median, mode, standard deviation)
    - Hypothesis testing
    - Confidence intervals
    - Regression analysis

Data Collection and Visualization

Effective data handling, processing, and visualization are critical for preparing data for analysis and communicating results. This involves cleaning, transforming, exploring, and visualizing data.

Estimated Time: 1–2 months

Essential Concepts

- **Data Cleaning**
    - Handling missing values
    - Removing duplicates •Outlier detection and treatment

- **Data Transformation**
    - Normalization and standardization
    - Encoding categorical variables
    - Feature scaling

- **Exploratory Data Analysis (EDA)**
    - Summary statistics
    - Data visualization (using libraries like Matplotlib, Seaborn)
    - Identifying patterns and correlations

- **Data Integration**
    - Merging and joining datasets
    - Data aggregation
    - Handling different data formats (CSV, JSON, SQL)

Machine Learning Fundamentals

Understanding machine learning fundamentals is crucial for building predictive models. This involves learning about different algorithms and how to train and evaluate models.

Estimated Time: 2-3 months

Essential Concepts

- **Supervised Learning**
  - Regression algorithms (e.g., linear regression, logistic regression)
  - Classification algorithms (e.g., decision trees, k-nearest neighbors, support vector machines)
- **Unsupervised Learning**
  - Clustering algorithms (e.g., K-means, hierarchical clustering)
  - Dimensionality reduction techniques (e.g., PCA, LDA
- **Model Evaluation**
  - Accuracy
  - Precision-Recall
  - F1 score
  - ROC-AUC
  - Confusion matrix
- **Model Training**
  - Train-test split
  - Cross-validation
  - Hyperparameter tuning
- **Overfitting and Underfitting**

•Recognizing overfitting and underfitting

•Techniques to mitigate overfitting (e.g., regularization, dropout)

•Model complexity management

Deep Learning

Deep learning is a subset of machine learning that involves neural networks with many layers. These models are powerful for handling large-scale data and complex patterns.

Estimated Time: 2▯3 months

Essential Concepts

- **Neural Networks**

  •Basics of neural networks

  •Activation functions •Forward and

  backward propagation

- **Advanced Neural Networks**

  •Convolutional Neural Networks (CNNs)

  •Recurrent Neural Networks (RNNs)

- **Deep Learning Frameworks**

  •Tools: TensorFlow, PyTorch, Keras

Specialization

Specializing in a specific area of data science allows you to develop expertise and stand out in the field. Two popular tracks are Natural Language Processing (NLP) and Computer Vision.

Estimated Time: 2□3 months

Essential Concepts

- **Natural Language Processing)NLP)**

  •Text preprocessing (tokenization, stemming, lemmatization)

  •Sentiment analysis •Named entity recognition □NER□ •Language modeling (using libraries like NLTK, SpaCy, Hugging Face)

- **Computer Vision**

  •Image Classification: Techniques and models

  •Object Detection: Algorithms like YOLO, SSD

  •Image Segmentation: Semantic and instance segmentation

  •Generative Models: GANs in computer vision

Big Data (Optional)

Big data skills are valuable for processing and analyzing large datasets, which is essential for certain data science roles. Understanding big data technologies can enhance your capabilities and make you more competitive in the job market.

Estimated Time: 2-3 months

Essential Concepts

- **Big Data Frameworks:** Hadoop, Spark
- **Data Processing:** MapReduce, Spark SQL
- **Data Storage:** HDFS, NoSQL databases (Cassandra, MongoDB)
- **Data Ingestion:** Kafka, Flume

Learning to code is a journey. Be patient with yourself and stay persistent, even when things get tough.

- Mosh