DATA SCIENTIST
CORE ROADMAP

A Step-by-Step Guide to Becoming a Data Scientist

6-8 Months

8 Phases

50+ Skills

Become a job-ready Data Scientist	Python, SQL, Sklearn, TF/PyTorch, MLflow	6–9 months (part-time)	ML Engineer / DS roles at top companies

Roadmap Overview

Phase 1	**Foundations of Mathematics & Statistics**	~4 Weeks
Phase 2	**Programming & Data Manipulation**	~4 Weeks
Phase 3	**Data Visualization & Storytelling**	~3 Weeks
Phase 4	**SQL & Databases**	~3 Weeks
Phase 5	**Machine Learning — Core Algorithms**	~8 Weeks
Phase 6	**Deep Learning & Advanced Topics**	~6 Weeks
Phase 7	**MLOps & Production Deployment**	~4 Weeks
Phase 8	**Portfolio, Projects & Job Readiness**	~4 Weeks

1 Foundations of Mathematics & Statistics

Weeks 1–4 | ~4 Weeks

PHASE 1

Every data scientist must master the mathematical backbone. This phase builds the quantitative intuition needed to understand algorithms, models, and data behavior.

Linear Algebra

- Vectors, matrices, and operations
- Matrix multiplication & transpose
- Eigenvalues & eigenvectors
- Dot product & norms
- Applications in ML (PCA, SVD)

Resources: Khan Academy Linear Algebra, Gilbert Strang MIT 18.06

Calculus

- Derivatives and partial derivatives
- Chain rule & product rule
- Gradient descent intuition
- Optimization (minima/maxima)
- Multivariable calculus basics

Resources: 3Blue1Brown Essence of Calculus, Paul's Online Math Notes

Statistics & Probability

- Descriptive stats: mean, median, mode, std
- Probability distributions (Normal, Binomial, Poisson)
- Bayes' Theorem
- Hypothesis testing & p-values
- Confidence intervals

Resources: StatQuest with Josh Starmer, Think Stats (free book)

Skills You'll Gain

Vectors	Matrices	Eigenvalues	Derivatives
Gradients	Probability	Distributions	Bayes
Hypothesis Testing	Confidence Intervals		

Phase Outcome: You can read research papers and understand algorithm derivations.

2 Programming & Data Manipulation

Weeks 5–8 | ~4 Weeks

PHASE 2

Python is the lingua franca of data science. Master the tools that let you collect, clean, and explore data at scale — the foundation of every real-world project.

Python Core

- Data types, loops, functions, OOP
- File I/O and error handling
- List comprehensions & generators
- Lambda, map, filter
- Virtual environments & pip
  - *Resources: Python.org docs, Automate the Boring Stuff (free)*

NumPy & Pandas

- NumPy arrays, broadcasting, indexing
- Pandas DataFrames — read, write, filter
- GroupBy, merge, pivot tables
- Handling missing values
- Time-series indexing
  - *Resources: Kaggle Learn Pandas, Official NumPy docs*

Data Collection & Cleaning

- Reading CSV, JSON, Excel, SQL
- Web scraping with BeautifulSoup/requests
- Data cleaning: duplicates, outliers, encoding
- Feature engineering basics
- EDA checklist
  - *Resources: Towards Data Science EDA guides, Kaggle notebooks*

Skills You'll Gain

Python	NumPy	Pandas	Data Cleaning
EDA	Web Scraping	Feature Engineering	GroupBy
Merging	Time-Series		

- **Phase Outcome: You can ingest any raw dataset and produce a clean, analysis-ready table.**

3 Data Visualization & Storytelling

Weeks 9–11 | ~3 Weeks

PHASE 3

Data without communication is noise. This phase teaches you to turn numbers into decisions using compelling charts, dashboards, and narratives.

Python Visualization

- Matplotlib: figures, axes, subplots
- Seaborn: statistical plots (heatmap, pairplot, violin)
- Plotly: interactive charts
- Choosing the right chart type
- Color theory and accessibility
  - *Resources: Matplotlib docs, Seaborn gallery, Plotly tutorials*

BI & Dashboards

- Tableau / Power BI basics
- Connecting to data sources
- Building interactive dashboards
- KPIs and metrics design
- Sharing and publishing reports
  - *Resources: Tableau Public free, Microsoft Power BI free tier*

Storytelling with Data

- Structure: context ➔ conflict ➔ resolution
- Annotation and highlighting
- Executive summaries
- Avoiding misleading charts
- Presenting to non-technical audiences
  - *Resources: Book: Storytelling with Data by Cole Nussbaumer*

Skills You'll Gain

Matplotlib	Seaborn	Plotly	Tableau
Power BI	Dashboard Design	Annotation	Storytelling
KPI Design	Chart Selection		

Phase Outcome: You can build a dashboard and present data insights to any stakeholder.

4 SQL & Databases

Weeks 12–14 | ~3 Weeks

PHASE 4

Most business data lives in relational databases. SQL mastery is non-negotiable — it's how you extract, join, and aggregate data before any model is trained.

SQL Fundamentals

- SELECT, WHERE, ORDER BY, LIMIT
- JOINs: INNER, LEFT, RIGHT, FULL
- GROUP BY, HAVING, aggregate functions
- Subqueries & CTEs
- CASE statements & conditional logic
  - *Resources: Mode Analytics SQL Tutorial, SQLZoo, W3Schools SQL*

Advanced SQL

- Window functions: ROW_NUMBER, RANK, LAG, LEAD
- Indexing and query optimization
- Stored procedures & views
- NULL handling
- Date/time functions
  - *Resources: LeetCode SQL 50, StrataScratch, DataLemur*

NoSQL & Cloud DBs

- MongoDB basics (documents, collections)
- When to use SQL vs NoSQL
- BigQuery / Snowflake intro
- Connecting Python to databases (SQLAlchemy, psycopg2)
- Data warehousing concepts
  - *Resources: MongoDB University, Google BigQuery sandbox (free)*

Skills You'll Gain

SQL	JOINs	CTEs	Window Functions
Subqueries	Indexing	BigQuery	MongoDB
SQLAlchemy	Data Warehousing		

- **Phase Outcome: You can write production-quality SQL queries to answer any business question.**

5 Machine Learning — Core Algorithms

Weeks 15–22 | ~8 Weeks

PHASE 5

This is the heart of data science. You'll learn when and why to apply each algorithm, not just how — building both intuition and implementation skills.

Supervised Learning

- Linear & Logistic Regression
- Decision Trees & Random Forests
- Gradient Boosting (XGBoost, LightGBM)
- Support Vector Machines
- K-Nearest Neighbors
- Naive Bayes
  - *Resources: Scikit-learn docs, Hands-On ML (Geron), StatQuest*

Unsupervised Learning

- K-Means & Hierarchical Clustering
- DBSCAN
- Principal Component Analysis (PCA)
- t-SNE & UMAP for visualization
- Anomaly detection
  - *Resources: Scikit-learn clustering guide, Kaggle competitions*

Model Evaluation & Tuning

- Train/Validation/Test split
- Cross-validation (k-fold, stratified)
- Metrics: accuracy, precision, recall, F1, AUC-ROC
- Hyperparameter tuning (GridSearchCV, Optuna)
- Bias-variance tradeoff
- SHAP for model explainability
  - *Resources: Scikit-learn model_selection, SHAP library docs*
  - **Skills You'll Gain**

Linear Regression	Logistic Regression	Random Forest	XGBoost
SVM	KMeans	PCA	Cross-Validation
AUC-ROC	SHAP	Optuna	Scikit-learn

Phase Outcome: You can train, evaluate, and explain ML models for real business problems.

6 Deep Learning & Advanced Topics

Weeks 23–28 | ~6 Weeks

PHASE 6

Modern AI is powered by neural networks. This phase equips you with deep learning skills that open doors to NLP, computer vision, and generative AI.

Neural Networks

- Perceptrons, activation functions, backpropagation
- Feedforward networks with Keras/TensorFlow
- PyTorch basics
- Regularization: dropout, batch norm
- Learning rate schedules
  - *Resources: fast.ai (free), Deep Learning Specialization (Coursera)*

Specialized Architectures

- CNNs for image classification
- RNNs & LSTMs for sequences
- Transformers & attention mechanism
- Transfer learning with pretrained models
- Fine-tuning BERT/GPT
  - *Resources: Hugging Face course (free), PyTorch tutorials*

NLP & LLM Applications

- Text preprocessing: tokenization, stemming
- TF-IDF & word embeddings (Word2Vec, GloVe)
- Sentiment analysis, NER, text classification
- Prompt engineering
- LangChain & RAG basics
  - *Resources: Hugging Face NLP course, LangChain docs*

Skills You'll Gain

TensorFlow	Keras	PyTorch	CNNs
LSTMs	Transformers	BERT	Transfer Learning
NLP	Embeddings	LangChain	Prompt Engineering

Phase Outcome: You can build image classifiers, text analyzers, and LLM-powered apps.

7 MLOps & Production Deployment

Weeks 29–32 | ~4 Weeks

PHASE 7

A model in a notebook is not a product. MLOps bridges the gap between experimentation and real-world impact — the skill that separates senior data scientists.

Version Control & Experiment Tracking

- Git & GitHub workflow for ML projects
- DVC for data version control
- MLflow experiment tracking
- Weights & Biases (W&B;)
- Reproducible pipelines
  - *Resources: MLflow docs, DVC.org, W&B; quickstart*

Model Deployment

- REST APIs with FastAPI / Flask
- Containerization with Docker
- Cloud deployment: AWS SageMaker, GCP Vertex AI, Azure ML
- Streamlit for ML demos
- Batch vs real-time inference
  - *Resources: FastAPI docs, Docker getting started, Streamlit docs*

Monitoring & CI/CD

- Data drift detection (Evidently AI)
- Model performance monitoring
- CI/CD pipelines for ML (GitHub Actions)
- Feature stores (Feast)
- A/B testing models in production
  - *Resources: Evidently AI docs, GitHub Actions for ML*

Skills You'll Gain

Git	DVC	MLflow	W&B;
FastAPI	Docker	AWS SageMaker	Streamlit
Evidently AI	CI/CD	Feature Store	A/B Testing

Phase Outcome: You can ship, monitor, and maintain ML models in production environments.

8 Portfolio, Projects & Job Readiness

Weeks 33–36 | ~4 Weeks

PHASE 8

Skills without evidence are invisible. This phase is about packaging your expertise into a portfolio that lands interviews — and preparing to ace them.

Portfolio Projects (Must Build)

- End-to-end ML project: data ➔ model ➔ API ➔ dashboard
- NLP project: sentiment analysis or chatbot
- Time-series forecasting (e.g. sales, stock)
- Kaggle competition top 20% finish
- Open-source contribution or blog post
  - *Resources: GitHub, Kaggle, Medium, Towards Data Science*

Resume & LinkedIn

- Quantify impact: 'Improved accuracy by 12%'
- Tailor resume per job description (ATS-friendly)
- LinkedIn: headline, about section, featured projects
- Get 3 recommendations
- Build a personal website or portfolio page
  - *Resources: Resume templates on Overleaf, Canva*

Interview Preparation

- Statistics & probability interview Qs
- ML fundamentals: algorithm trade-offs
- SQL interview (LeetCode, DataLemur)
- Case studies: A/B testing, metrics design
- System design: ML system at scale
- Behavioral (STAR method)
  - *Resources: LeetCode SQL, DataLemur, Ace the Data Science Interview book*

Skills You'll Gain

Portfolio	GitHub	Kaggle	Storytelling
Resume	LinkedIn	SQL Interviews	ML Interviews
System Design	Case Studies	A/B Testing	Communication

Phase Outcome: You are ready to apply, interview, and land your first Data Scientist role.

Pro Tips for Success

Consistency Over Intensity	2 hours daily beats 14 hours on weekends. Build the habit, not the sprint.
Document Everything	Keep a learning journal. Write your own notes in plain English — feynman technique.
Build Projects Early	Start building from Phase 2. Imperfect projects > perfect notebooks.
Join the Community	Kaggle, Reddit r/datascience, LinkedIn — share your work, get feedback.
Understand, Don't Memorize	Know WHY each algorithm works. You'll forget syntax, but intuition lasts.
Kaggle is your Gym	Participate in competitions. Even finishing last teaches you more than tutorials.