Mastering Data Science: From Scratch to a World-Class Level

This roadmap provides a comprehensive guide to mastering data science, integrating artificial intelligence (AI), machine learning (ML), deep learning (DL), large language models (LLMs), information retrieval (IR), recommender systems, agentic AI, and the latest advancements in AI infrastructure, including vector databases and retrieval-augmented generation (RAG).

---

1. Foundational Knowledge (Weeks 1-4)

Mathematics & Statistics

- Linear Algebra: Matrix operations, eigenvalues/eigenvectors, SVD.
- Calculus: Differentiation, optimization techniques.
- Probability & Statistics: Bayesian inference, hypothesis testing, confidence intervals.

Programming & Software Engineering

- Python (NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn).
- Best practices: Git, version control, testing frameworks.
- Shell scripting and automation.

Databases & Data Storage

- SQL: PostgreSQL, MySQL, indexing, query optimization.
- NoSQL: MongoDB, Redis, Elasticsearch for unstructured data.
- Cloud storage: AWS S3, Google Cloud Storage, Azure Blob Storage.

---

2. Data Wrangling, Preprocessing & Visualization (Weeks 5-8)

- Handling missing data, outliers, feature scaling.
- Feature engineering for ML and DL.
- Advanced visualization: Interactive dashboards (Plotly, Power BI, Tableau).

---

3. Machine Learning (Weeks 9-20)

Supervised Learning

- Regression (Linear, Ridge, Lasso, ElasticNet).
- Classification (Logistic Regression, SVM, Decision Trees, Random Forest, XGBoost, LightGBM, CatBoost).

Unsupervised Learning

- Clustering (K-means, DBSCAN, Hierarchical Clustering).
- Dimensionality Reduction (PCA, t-SNE, UMAP).

Time Series Analysis

- ARIMA, SARIMA, Prophet for forecasting.
- LSTMs and Transformers for time series prediction.

Model Optimization & Deployment

- Hyperparameter tuning (Grid Search, Bayesian Optimization, Optuna).
- Model Explainability (SHAP, LIME, interpretability techniques).
- Model deployment (Flask, FastAPI, Docker, Kubernetes).

---

4. Deep Learning & Large-Scale AI Models (Weeks 21-30)

Deep Learning Foundations

- Neural Networks (MLPs, Backpropagation, Optimizers like Adam, SGD).
- Frameworks: TensorFlow, PyTorch, JAX.

Computer Vision

- Convolutional Neural Networks (CNNs: ResNet, EfficientNet, Vision Transformers).
- Image segmentation (U-Net, Mask R-CNN).
- Generative models (GANs, Stable Diffusion, StyleGAN).

Natural Language Processing (NLP)

- Word Embeddings (Word2Vec, GloVe, FastText).
- Transformer-based models (BERT, GPT-4, T5, LLaMA, Falcon).
- Fine-tuning & prompt engineering for LLMs.

5. Advanced AI Architectures & Vector Databases (Weeks 31-40)

Information Retrieval (IR) & Search Systems

- TF-IDF, BM25, Elasticsearch for keyword-based search.
- Neural retrieval models (ColBERT, DPR, RAG).

Vector Databases & Retrieval-Augmented Generation (RAG)

- Introduction to vector embeddings (FAISS, Annoy, ScaNN).
- Using vector databases (Pinecone, Weaviate, ChromaDB, Milvus) for real-time search.
- Combining LLMs with vector search for context-aware AI systems.

Multimodal AI

- Combining vision, text, and speech (CLIP, DALL·E, Whisper).
- Applications in AI-generated art, video analysis, speech recognition.

6. Recommender Systems (Weeks 41-50)

- **Collaborative Filtering:** Matrix Factorization, SVD, ALS.
- **Content-based Filtering:** TF-IDF, word embeddings for recommendations.
- **Hybrid Systems:** Combining collaborative and content-based filtering.
- **Real-time Personalization:** Context-aware and reinforcement learning-based recommender systems.

7. Reinforcement Learning & Agentic AI (Weeks 51-60)

- **Markov Decision Processes (MDPs), Dynamic Programming.**
- **Deep RL:** DQN, PPO, SAC, MuZero.
- **Autonomous AI Agents:** AI planning, self-learning agents, multi-agent reinforcement learning (MARL).
- **Applications:** Game AI, robotics, self-driving cars, real-time trading systems.

8. Industry Applications & Scalability (Weeks 61-75)

AI for Healthcare

- Medical imaging (AI-assisted diagnosis, CT/MRI analysis).
- Predictive analytics for patient risk assessment.

AI for Finance & Business Intelligence

- Algorithmic trading, fraud detection, risk modeling.
- Customer analytics and churn prediction.

Enterprise AI Deployment

- Edge AI: Running models on mobile devices, IoT (TensorFlow Lite, ONNX).
- Serverless AI: AWS Lambda, Google Cloud Functions.
- MLOps: CI/CD pipelines for AI (MLflow, DVC, Kubeflow).

---

9. Staying on the Cutting Edge & Continuous Learning

- **AI Research:** Reading latest papers (ArXiv, NeurIPS, ICML, CVPR, ACL).
- **Kaggle & Hackathons:** Competing in AI/ML challenges.
- **Networking & Open Source:** Contributing to AI projects on GitHub.

---

10. Final Mastery & Integration of Skills

- **Building AI-first products:** Deploy LLM-powered chatbots, intelligent search engines, multimodal AI systems.
- **Startup & Entrepreneurship:** AI-powered SaaS products, enterprise AI solutions.
- **Research & Innovation:** Pushing boundaries in AI safety, ethical AI, and autonomous AI.