Introduction to Data Science

Module 1.1

Sunit Bhattacharya

July 2025

Why Data Science?

Module 1.1: The Foundations and Impact of Data
Module 1.2: A Pythonic Introduction to Data Science

Why Data Science Matters

- **Data is the new oil:** Driving progress in science, industry, health, and policy.
- **Data = Decisions:** From personalized medicine to targeted advertising.
- **Career Impact:** Top roles in tech, finance, research, and startups.
- **Scale of Impact:**
  - AI boom driven by data and GPUs.
  - NVIDIA hit $4T market cap[1] through AI/data acceleration.
- **Analogy:** Just as electricity powered the Industrial Revolution, data fuels today's digital transformation.

[1] As of July 2025.

NVIDIA: A Data Science Powerhouse Transforming the Future

- NVIDIA's GPUs have become the **engine driving the global AI revolution**, powering breakthroughs from conversational AI to self-driving cars.
- Their astounding market value surge reflects:
  - **Explosive growth** in AI applications demanding immense computational power.
  - Pioneering innovations in *deep learning* and generative AI models that seem like science fiction come to life.
  - A carefully crafted ecosystem that **accelerates data science workflows**—making complex computations faster and more accessible worldwide.
- **Why it matters:** NVIDIA's story shows how data science shapes modern companies, economies, and our everyday lives.
- *Imagine*: From video games to the AI assistants you use, NVIDIA's technology is at the heart of this new data-driven epoch.

NVIDIA: Case Study in Data Science Impact

Economic and Market Value Comparison (July 2025)

Entity	Value	Core Driver
NVIDIA	**$4.07T**	AI/data science infrastructure
Microsoft	**$3.75T**	Cloud & AI apps
Apple	**$3T**	Consumer tech, AI
India (Est. GDP)	**$4.7T**	Domestic demand, services, agriculture
Japan (Est. GDP)	**$4.19T**	Technology, manufacturing, services

Note: Market caps are approximate as of July 2025. GDP figures are 2025 projections (T=Trillion).

Career Avenues in Data Science

- **Data Scientist:** Builds models, analyzes complex datasets, and derives actionable insights using statistics, machine learning, and programming.
- **Data Analyst:** Interprets data, generates reports, and visualizes trends to guide decision-making.
- **Machine Learning Engineer:** Designs, develops, and deploys ML models for prediction and automation.
- **Data Engineer:** Builds and maintains data pipelines and infrastructure to enable robust data flow and large-scale analytics.
- **Business Intelligence (BI) Analyst/Developer:** Develops dashboards, conducts reporting, and drives business decisions with analytics tools.
- **AI/ML Research Scientist:** Advances new algorithms and techniques in artificial intelligence and machine learning.

Career Avenues in Data Science

- **Data Architect:** Designs, structures, and manages large-scale data systems for scalability and performance.
- **Statistician:** Applies advanced statistical methods to analyze data and extract insights.
- **NLP Specialist:** Develops solutions to analyze text, speech, and language data for applications like chatbots, sentiment analysis, and information retrieval.
- **Big Data Engineer:** Works with massive datasets using distributed systems (Hadoop, Spark) to enable processing and analysis at scale.
- **Product/Data Strategy Analyst:** Integrates analytics with product and business strategy to guide innovation and growth.
- **Data & Analytics Manager/Leader:** Oversees teams, drives analytics strategy, and ensures alignment with organizational goals.

Career Avenues in Data Science for Economics Majors in India

- **Data Analyst / Research Associate:** Analyze economic datasets, survey data, and government statistics (e.g. RBI reports) to inform research and policy studies.
- **Economic Data Scientist:** Build predictive models for financial analysis, policy impact evaluation, and market research.
- **Quantitative Researcher for Financial Institutions:** Develop econometric and machine learning models for risk management, algorithmic trading, and credit scoring in banks and fintech.
- **Policy Analyst / Data Specialist:** Support government and think tanks (e.g., NITI Aayog) using data-driven insights to design and evaluate socio-economic programs.
- **Business Intelligence and Analytics Consultant:** Work with Indian industry sectors (insurance, telecom, retail) designing dashboards, KPIs, and strategic analytics for business growth.

Career Avenues in Data Science for Economics Majors in India

- **Data Engineer / Infrastructure Specialist:** Maintain and optimize data pipelines with an emphasis on large Indian government and enterprise datasets.
- **ML Engineer / AI Specialist in Economic Domains:** Implement models for credit risk, fraud detection, customer analytics in Indian banks, NBFCs, and digital platforms.
- **Academia and Research Scientist:** Conduct interdisciplinary research linking econometrics, AI, and data science in Indian universities or research institutes.
- **Data & Analytics Manager / Chief Data Officer:** Lead analytics teams in Indian enterprises or government agencies aligning data strategy with economic development goals.

Career Avenues in Data Science

Career Progression: Entry-level **(Research Associate / Data Analyst)** → Mid-level **(Economic Data Scientist / BI Consultant / ML Engineer)** → Senior/Leadership **(Policy Advisor / Analytics Manager / Chief Data Officer)**

What is Data Science?

- **Data Science** is a dynamic, interdisciplinary field that combines **computer science**, **statistics**, and **domain expertise** to uncover meaningful insights hidden within data.
- At its core, data science is about turning raw data into knowledge that drives smarter decisions and impactful actions.
- The **four pillars** of data science:
  - **Curiosity & Questions:** Defining the right problems to solve.
  - **Data Wrangling:** Collecting, cleaning, and preparing complex, often messy data.
  - **Analysis & Modeling:** Finding patterns, building predictive models, and validating results.
  - **Communication:** Translating discoveries into clear, actionable insights for decision-makers.
- By integrating **statistics**, **computer science**, **cognitive science**, and real-world knowledge, data science empowers us to understand the world in new and powerful ways.

What is Data Science?

- **Core Idea:** Data science is the process of converting **raw, noisy, high-dimensional data** into **actionable insight and knowledge**.
- **Mathematical Framing:**
  - Learn a function $f : \mathcal{X} \to \mathcal{Y}$ such that:
    - $\mathcal{X}$ = feature/input space (e.g., data about individuals, markets, sensors)
    - $\mathcal{Y}$ = target/output space (e.g., categories, real values, actions)
  - The function f is **estimated from data**, often using statistical, algorithmic, or machine learning methods.
- **Interdisciplinary Nature:**
  - Combines **statistics**, **computer science**, **optimization**, and **domain expertise**.
  - Tools: inference, prediction, modeling, visualization, automation.

What is Data Science?

Econometrics vs Data Science:

- Econometrics: Often focused on *causal inference* with structured/tabular data.
- Data science: Broader scope—handles unstructured data (text, images), real-time streams, and emphasizes *prediction and scalability*.
- Both rely on rigorous modeling, but with different emphases and assumptions.

Real life analogy

> *"Econometrics helps policymakers evaluate past interventions. Data science helps them decide the next best action using all available data."*

What Counts as Data?

Data Types:

- **Structured:** Tables, spreadsheets, SQL databases
- **Unstructured:** Text documents, images, audio files
- **Semi-structured:** JSON, XML, API responses
- **Streaming:** Financial tick data, sensor logs, real-time feeds

Data Sources (Examples):

- Government Sources: Open Government Data Platform
- University Repos: NYU
- Kaggle datasets: US Funds Dataset
- public APIs: MarketStack

Core Tools of the Data Scientist

- **Languages:** Python (Pandas, NumPy, Scikit-learn), R
- **Data Extraction**: POSTMAN
- **Workflow:** Jupyter, GitHub, Colab, VSCode
- **Libraries:**
    - Data: Pandas, NumPy
    - Visualization: Matplotlib, Seaborn
    - Machine Learning: Scikit-learn, XGBoost
    - Deep Learning: PyTorch
- **AI Tools:** ChatGPT, Perplexity, Gemini, Cursor
- **Formats:** CSV, JSON, Parquet, APIs

The Data Science Lifecycle: A Roadmap for Insight

1. **Problem Definition & Understanding**
   Clarify the question and align goals with stakeholders
2. **Data Collection**
   Gather relevant data from diverse sources
3. **Data Preparation**
   Clean, integrate, and transform raw data into usable form
4. **Exploratory Data Analysis (EDA)**
   Discover patterns, visualize trends, and generate hypotheses
5. **Modeling & Algorithm Selection**
   Build and tune predictive or descriptive models
6. **Evaluation & Validation**
   Assess model performance and generalizability
7. **Deployment**
   Integrate insights and models into business workflows
8. **Monitoring & Maintenance**
   Track performance, update models, and ensure lasting value

Remember: This lifecycle is iterative — re-visit and refine.

Lifecycle: Initial Phases

Problem Definition: Clearly understand the business or research goal. Engage stakeholders to align expectations and define success criteria.

Data Collection: Gather data from relevant sources such as databases, APIs, sensors, and external data providers. Record data provenance and ensure data access permissions.

Preparation: Clean and transform the raw data. Handle missing or inconsistent values, correct errors, and integrate heterogeneous data sources to prepare for analysis.

Exploratory Data Analysis (EDA): Use statistical summaries, visualizations, and profiling techniques to uncover patterns, identify outliers, detect biases, and understand data distributions.

Lifecycle: Modeling to Monitoring

Modeling: Select and build appropriate models such as regression, classification, or clustering. Train models on data and tune hyperparameters to optimize performance.

Evaluation: Assess model effectiveness using relevant metrics (e.g., accuracy, RMSE, recall). Perform validation and ensure robustness before deployment.

Deployment: Integrate the finalized model into business workflows via APIs, dashboards, or applications. Ensure scalability and security compliance.

Monitoring: Continuously track model performance to detect drift or degradation. Update or retrain models as new data, trends, or concepts evolve to maintain effectiveness.

Modeling: From Statistics to Machine Learning

- **Supervised Learning:** Models trained on labeled data, where the goal is to predict an output variable based on input features. Examples include *regression* for continuous outcomes (e.g., predicting house prices) and *classification* for categorical outcomes (e.g., spam detection). These methods often rely on assumptions about data distribution and noise.
- **Unsupervised Learning:** Models identify structure or patterns in data without predefined labels. Common techniques include *clustering* (grouping similar data points) and *principal component analysis (PCA)* for dimensionality reduction and visualization. These help in feature extraction and discovering hidden insights.

Modeling: From Statistics to Machine Learning

- **Common Algorithms:** From classical statistics to modern ML, key algorithms include:
  - *Logistic Regression*: a probabilistic model for binary classification.
  - *Decision Trees*: hierarchical models splitting data based on feature thresholds.
  - *Support Vector Machines (SVM)*: maximizing class separation using hyperplanes.
  - *Neural Networks*: layers of interconnected nodes for learning complex, non-linear patterns.
- **Bridging Econometrics and ML:** Regression trees can be seen as a flexible extension of linear regression—splitting data space into regions where simple linear models apply (piecewise-linear). This illustrates how ML models generalize traditional statistical tools to capture complex relationships adaptively, without relying solely on parametric assumptions.

Example-Driven Introduction: A Toy Problem

Project: Predicting Student Dropout Risk

- **Goal:** Build a model to identify university students at high risk of dropping out.
- **Features:**
  - GPA, Attendance(%), Major, OnCampusResident, FinancialAid, ClubsInvolved, PreviousWarnings, CreditsCompleted, Age, etc.
- **Target:** Dropout ($1 = $ Yes, $0 = $ No)

Why is this "hard"?

- Multiple, messy, and interacting variables.
- Incomplete data; imbalanced outcome.
- Real-world, high-stakes consequences for false positives/negatives.

How This Problem Illustrates the Lifecycle

1. **Problem Definition:** Identify “at risk” students.
2. **Data Collection:** Student academic/personnel records.
3. **Preparation:** Cleaning missing club data, standardizing majors.
4. **EDA:** Visualize GPA distributions, dropout rates by major.
5. **Modeling:** Logistic regression or decision tree classifier.
6. **Evaluation:** Test on unseen data; discuss trade-off (precision/recall).
7. **Deployment:** Alerts/support to university advisors.
8. **Monitoring:** Update as student profiles/patterns change.

Key Takeaways

- Data science transforms industries, opportunities, and individual careers.
- Core workflow (lifecyle) applies across domains—from detecting dropouts to powering trillion-dollar tech companies.
- Example-driven learning demystifies complexity and enables real-world understanding.

Next: Dive into core Python tools and your first hands-on data exploration!