OXFORD ARTIFICIAL INTELLIGENCE SOCIETY

An introduction to Data Science

“Data science is the process of using algorithms, methods, and systems to extract knowledge and insights from structured and unstructured data. It uses analytics and machine learning to help users make predictions, enhance optimization, and improve operations and decision making.” – IBM

Today’s data science teams are expected to answer many questions. Business demands better prediction and optimization based on real-time insights backed by tools like these.

The data science lifecycle starts with gathering data from relevant sources, cleaning it and putting it in formats that can be understood by computers. In the next phase, statistical methods and other algorithms are used to find patterns and trends. Then models are programmed and built to predict and forecast and finally, results are interpreted.

Advances in AI, machine learning and automation have raised the standards of data science tools for business. The result is the formation of data science teams, expert data scientists, citizen data scientists, programmers, engineers and business analysts, that extend across business units.

The opportunity here is massive. The automation of tedious data science tasks such as data preparation, and the empowerment of analysts to build models, keeps business agile and innovative. Automating the data science lifecycle frees expert data scientists to address the more interesting and innovative aspects of the field. Human intelligence, combined with data science technology and automation, helps a business extract greater value from data.

With the volume and variety of social, mobile and device data, along with new technologies and tools, data science today plays a broader role than ever before. Business considers data science and AI to be a technology-enabled strategy. In order for data science to be effective, its full lifecycle not only must support traditional analytics, but it must also work in concert with modern applications. This means that the data science practice must evolve beyond routine, tedious tasks — as much 85% of a data scientist’s time is spent cleaning, shaping and moving data from place to place, often to feed machine learning. That leaves only a small percentage of time to find patterns and trends, to build models, to predict and forecast, and to interpret results.

Fortunately, there is relief. The latest development in modern data science is an AutoAI capability that automates the data preparation and modeling stages of the data science lifecycle. Now, not only can more data scientists use their specialized skills the way they were intended; but more businesses can benefit from data science, from prediction to automisation.

Pros of being a data scientist

- It’s in Demand: Data Science is greatly in demand. Prospective job seekers have numerous opportunities. It is the fastest growing job on Linkedin and is predicted to create 11.5 million jobs by 2026.
- Abundance of Positions: There are very few people who have the required skill-set to become a complete Data Scientist. This makes Data Science less saturated as compared with other IT sectors.
- Data Science is Versatile: Data Science is a very versatile field, and you will have the opportunity to work in various fields.

Cons of being a data scientist

- Data Science is Blurry Term: Data Science is a very general term and does not have a definite definition. A Data Scientist’s specific role depends on the field that the company is specializing in.
- Mastering Data Science is near to impossible: Being a mixture of many fields, Data Science stems from Statistics, Computer Science and Mathematics. It is far from possible to master each field and be equivalently expert in all of them. Therefore, it is an ever-changing, dynamic field that requires the person to keep learning the various avenues of Data Science.
- A large Amount of Domain Knowledge Required: A person without a considerable background in Statistics and Computer Science will find it difficult to solve Data Science problem without its background knowledge.

Types of Job

What does a data scientist do?

A data scientist works with data to draw out meaning and insightful conclusions that can drive decision making in an institution or organization.

Their job role includes data collection, data transformation, data visualization, and analysis, building predictive models, providing recommendations on actions to implement based on data findings.

Data scientists work in different sectors such as healthcare, government, industries, energy, academia, technology, entertainment, etc.

Some top companies that hire data scientists are Amazon, Google, Microsoft, Facebook, LinkedIn, Twitter, Netflix, IBM, etc.

1. Data Analyst

Data analysts are responsible for a variety of tasks including visualisation, munging, and processing of massive amounts of data. They also have to perform queries on the databases from time to time. One of the most important skills of a **data analyst** is optimization. This is because they have to create and modify algorithms that can be used to cull information from some of the biggest databases without corrupting the data.

How to Become a Data Analyst?

SQL, R, SAS, Python are some of the sought after technologies for **data analysis**. So, certification in these can easily give a boost to your job applications. You should also have good problem-solving qualities.

2. Data Engineers

Data engineers build and test scalable Big Data ecosystems for the businesses so that the **data scientists** can run their algorithms on the data systems that are stable and highly optimized. **Data engineers** also update the existing systems with newer or upgraded versions of the current technologies to improve the efficiency of the databases.

How to Become a Data Engineer?

If you are interested in a career as a **data engineer**, then technologies that require hands-on experience include Hive, NoSQL, R, Ruby, Java, C++, and Matlab. It would also help if you can work with popular data APIs and ETL tools, etc.

3. Database Administrator

The job profile of a database administrator is pretty much self-explanatory- they are responsible for the proper functioning of all the databases of an enterprise and grant or revoke its services to the employees of the company depending on their requirements. They are also responsible for database backups and recoveries.

How to Become a Database Administrator?

Some of the essential skills and talents of a database administrator include database backup and recovery, data security, data modelling, and design, etc. If you are good at disaster management, it’s certainly a bonus.

4. Machine Learning Engineer

Machine learning engineers are in high demand today. However, the job profile comes with its challenges. Apart from having in-depth knowledge in some of the most powerful technologies such as SQL, REST APIs, etc. machine learning engineers are also expected to perform A/B testing, build data pipelines, and implement common machine learning algorithms such as classification, clustering, etc.

How to Become a Machine Learning Engineer?

Firstly, you must have a sound knowledge of some of the technologies like Java, Python, JS, etc. Secondly, you should have a strong grasp of statistics and mathematics. Once you have mastered both, it’s a lot easier to crack a job interview.

5. Data Scientist

Data scientists have to understand the challenges of business and offer the best solutions using data analysis and data processing. For instance, they are expected to perform predictive analysis and run a fine-toothed comb through an “unstructured/disorganized” data to offer actionable insights. They can also do this by identifying trends and patterns that can help the companies in making better decisions.

How to Become a Data Scientist?
To become a data scientist, you have to be an expert in R, MatLab, SQL, Python, and other complementary technologies. It can also help if you have a higher degree in mathematics or computer engineering, etc.

6. Data Architect

A data architect creates the blueprints for data management so that the databases can be easily integrated, centralized, and protected with the best security measures. They also ensure that the data engineers have the best tools and systems to work with.

How to Become a Data Architect?
A career in data architecture requires expertise in data warehousing, data modelling, extraction transformation and loan (ETL), etc. You also must be well versed in Hive, Pig, and Spark, etc.

7. Statistician

A statistician, as the name suggests, has a sound understanding of statistical theories and data organization. Not only do they extract and offer valuable insights from the data clusters, but they also help create new methodologies for the engineers to apply.

How to Become a Statistician?
A statistician has to have a passion for logic. They are also good with a variety of database systems such as SQL, data mining, and the various machine learning technologies.

8. Business Analyst

The role of **business analysts** is slightly different than other data science jobs. While they do have a good understanding of how data-oriented technologies work and how to handle large volumes of data, they also separate the high-value data from the low-value data. In other words, they identify how the **Big Data** can be linked to actionable business insights for business growth.

How to Become a Business Analyst?
Business analysts act as a link between the data engineers and the management executives. So, they should have an understanding of business finances and **business intelligence**, and also the IT technologies like data modelling, data visualization tools, etc.

9. Data and Analytics Manager

A **data and analytics** manager oversees the data science operations and assigns the duties to their team according to skills and expertise. Their strengths should include technologies like SAS, R, SQL, etc. and of course management.

Skills and Experience

A data scientist is better statistician than any software engineer and better engineer as compared to any statistician. Data Scientist needs to have both technical and non-technical skills to perform their job in an effective manner.

Technical Skills

Statistical & Probability Skills

Statistical Thinking is the most important aspect of Data Science. Generally, Statistics is divided into two categories:

- Descriptive Statistics deals with summarizing and describing the data. It quantitatively summarizes large features of data through visualizations and outlines the sample from a larger population of data values.

- Inferential Statistics is about inferring or concluding from the data. It is about drawing conclusions from a smaller sample and implying the drawn conclusions over a larger group.

Another skill that is needed to become a data scientist is Probability. Concepts of probability is the backbone of data science and one must be skilled at it in order to carry out complex machine learning operations. It is also a key skill that helps you to ascertain the uncertainty and chances of events.

I recently read an article based on an interview with a machine learning researcher at Google Deepmind and he suggested reading recent publications and developing computer codes based on them. He said that doing so improves the programming skills as well as learn what recent trends are in the industry.

Mathematical Skills

Mathematics is another important part of Data Science. If you want to become a proficient Data Scientist, then you must be proficient in the topics below.

- Linear Algebra powers everything that runs on Machine Learning. It is used in the artistic rendering of your photographs, recommendation systems and facial recognition.

- Calculus is used extensively in Data Science, especially in calculating loss function which is the most important concept in optimizing models. The concept of partial derivates is also used in backpropagation for neural networks.

- Discrete Math is the study of values that are distinct and separate. The topics of Discrete Math include Boolean Algebra, Set Theory, Relations & Functions, number theory, recursion, and graph theory. Discrete Math is also useful when dealing with databases, for example, the set theory can be applied to the inner-joins and outer-joins of the table.

- Optimization Theory teaches you how to find the most optimal solution in a complex multi-dimensional space. It allows you to make the best out of data and develop better models. There are three parts of optimization: Variables, Constraints and Objective Function.

Programming Skills

Programming allows you to implement your statistical thinking in a practical setting.

- Python is the easiest programming language that you can learn for Data Science. Python is highly versatile, and comes with a wide range of libraries and functions that you can implement in your code to develop robust models. Some of the libraries that you must know for Data Science are Pandas (data wrangling), Matplotlib (data visualization), Numpy (matrix manipulation), Scikit-learn (machine learning), and TensorFlow (deep learning).

- R is a statistical programming tool that is used for solving core-data science problems. Some of the important packages of R that you must be skilled at in order to become data scientist are ggplot2 (data visualization), dplyr (data manipulation), purrr (data wrangling).

- Database Query Languages. A relational database is a collection of structured data in rows and columns. This form of data is usually generated by mobile devices, IoT devices, and services that can be easily managed. You must be skilled at SQL which is designed for querying database models, and NoSQL which allows you to deal with an unstructured form of data. Some of the SQL languages are MySQL, PL/SQL etc. whereas NoSQL languages are MongoDB, Cassandra, Redis, etc.

- Big Data Technologies. The knowledge of Big Data is highly treasured by the industries. Some of the trending big data technologies are Apache Hadoop (open source big data platform written in Java, and interoperable in multiple programming languages like Java, Python, C++, Perl, and Ruby) and Apache Spark (for real-time streaming management of data).

Non-Technical Skills

Data Inquisitiveness

Inquisitiveness or curiosity to learn more is the key towards acquiring mastery of any quantitative field. Since Data Science is highly quantitative in nature, and it is constantly evolving. Thus, you must stay ahead of the curve by updating yourself with articles, blogs, new updates in programming languages, tools, etc. This requires a high magnitude of intellectual curiosity for learning new concepts and implementing them.

Business Expertise

Data Science revolves around the business domain and therefore requires the data scientist to have knowledge of the business requirements. The main goal of a data scientist is to translate business problems into data science solutions through the implementation of analytical skills.

Communication Skills

Communication Skills are utmost important for Data Scientists. Some of the important areas in Data Science where communication skills are important as Data Visualization and Storytelling.

Teamwork

Data Scientists work on projects that require the combined efforts of several team members. As a Data Scientist, you have to work with several members of the company like business analysts for understanding customer requirements, marketing department and software team for product development.

Getting a Job

What is the job outlook for data scientists?

The job outlook for data scientists is very positive. IBM predicts the demand for data scientists to soar 28% by 2020. A recent study using the LinkedIn job search tool shows that a majority of top tech jobs in the year 2020 are jobs that require skills in data science, business analytics, machine learning, and cloud computing.

How much do data scientists make?

Data scientist salaries depend on the organization or company you are working for, your educational background, number of years of experience, and your specific job role.

Data scientists make anywhere from $50,000 to $250,000 with the median salary being about $120,000. This article discusses more the salaries of data scientists.

Resources for learning about data science

Generally, if you have a solid background in an analytic discipline such as physics, mathematics, economics, engineering, or computer science, and you are interested in exploring the field of data science, the best way is to begin with massive open online courses (MOOCs).

Then after establishing a solid foundation, you may then seek other ways to increase your knowledge and expertise such as studying from textbooks, engaging in projects, and networking with other data science aspirants.

Full list of technologies you might consider learning:

- Robotic Process Automation (RPA): technology that extracts the list of rules and actions to perform by watching the user doing a certain task
- Expert Systems: a computer program that has hard-coded rules to emulate the human decision-making process.
- Computer Vision (CV): methods to acquire and make sense of digital images; usually divided into activities recognition, images recognition, and machine vision.
- Natural Language Processing (NLP): sub-field that handles natural language data; usually divided into language understanding, language generation, and machine translation
- Neural Networks (NNs or ANNs): a class of algorithms loosely modelled after the neuronal structure of the human/animal brain that improves its performance without being explicitly instructed on how to do so. The two majors and well-known sub-classes of NNs are Deep Learning and Generative Adversarial Networks.
- Autonomous Systems: sub-field that lies at the intersection between robotics and intelligent systems; it deals with tasks such as intelligent perception, dexterous object manipulation, and plan-based robot control.
- Distributed Artificial Intelligence (DAI): a class of technologies that solve problems by distributing them to autonomous agents that interact with each other. Multi-agent

systems (MAS), Agent-based modelling (ABM), and Swarm Intelligence are three useful specifications of this subset, where collective behaviours emerge from the interaction of decentralized self-organized agents;

- Affective Computing: a sub-field that deal with emotions recognition, interpretation, and simulation;
- Evolutionary Algorithms (EA): it is a subset of a broader computer science domain called evolutionary computation that uses mechanisms inspired by biology to look for optimal solutions.
- Inductive Logic Programming (ILP): sub-field that uses formal logic to represent a database of facts and formulate hypothesis deriving from those data;
- Decision Networks: is a generalization of Bayesian networks/inference, which represent a set of variables and their probabilistic relationships through a map.
- Probabilistic Programming: a framework that does not force you to hardcode specific variable but rather works with probabilistic models.
- Ambient Intelligence (AmI): a framework that demands physical devices into digital environments to sense, perceive, and respond with context awareness to an external stimulus.

Reference links

- Jobs reference: https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/
- Introduction: https://www.ibm.com/uk-en/analytics/data-science
- Pros and cons: https://towardsdatascience.com/why-choose-data-science-for-your-career-ca38db0c28d4
- Skills: https://data-flair.training/blogs/skills-needed-to-become-a-data-scientist/
- Job outlook: https://medium.com/towards-artificial-intelligence/answers-to-questions-from-data-science-aspirants-f2950be3cd18
- External resources: https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020