Module 2. Data science: career opportunities in sport

Unit 2.1 Career opportunities in data science

In the last module, we explored the definition and history of data, how sports have been able to generate big data, and its profound impact on the sports industry. The rise of data science is particularly noticeable in team sports, where analytics departments have become integral to both sport/technical side and front-office decisions. For instance, the top 5 professional football leagues, such as the English Premier League, La Liga and others, have increasingly recognized the value of data scientists. They have placed full-functioning departments with divided roles and responsibilities to optimize performance, scout talent and strategize match plans.

Learning data science or programming follows the same trajectory as learning a new language. As speaking or writing skills improve with practice, so do coding and data science abilities. The more you code, the more fluent you become. Data science is the art of blending mathematics with critical thinking. As many spoken languages can be traced back to a common ancestor, programming languages share foundational similarities. For example, understanding the basics of languages like C or Java can help you learn Python and R easily. Learning a language relies heavily on practice.

“While the term data science is not new, the meanings and connotations have changed over time. The word first appeared in the ’60s as an alternative name for statistics. In the late ’90s, computer science professionals formalized the term. A proposed definition for data science saw it as a separate field with three aspects: data design, collection, and analysis. It still took another decade for the term to be used outside academia” (Amazon Q, n.d., https://lc.cx/nDijJG).

Over the past decade, the growth of sports data companies has reached a heightened demand to collect, process, and provide data. Data is generated from various aspects of sports, ranging from play-by-play statistics to the Olympians' real-time performance metrics. This data provides valuable insights into the different dimensions of sports performance, helping teams make informed decisions. As the volume of data generated by sports companies grows exponentially, the necessity to process, organize, and clean the data has guided the sports industry into a new era of roles and responsibilities. This intersection of new technology (mentioned in module 1.5) has led to a diversification in job opportunities among sports organizations.

2.1.1 What is the data science process?

“Once the problem has been defined, the data science team may solve it using the **OSEMN** data science process.

O – Obtain data

Data can be pre-existing or acquired for a data repository online. Data scientists can extract data from internal or external databases.

S – Scrub data

Data cleaning is the process of standardizing the data according to a predetermined format. It includes handling missing data, fixing errors, and removing data outliers.

E – Explore data

Data exploration is preliminary data analysis using descriptive statistics and data visualization tools. Then, they explore the data to identify interesting patterns that can be studied or actioned.

M – Model data

Software and machine learning algorithms are used to gain deeper insights, predict outcomes, and prescribe the best action. Machine learning techniques like association, classification, and clustering are applied to the training data set.

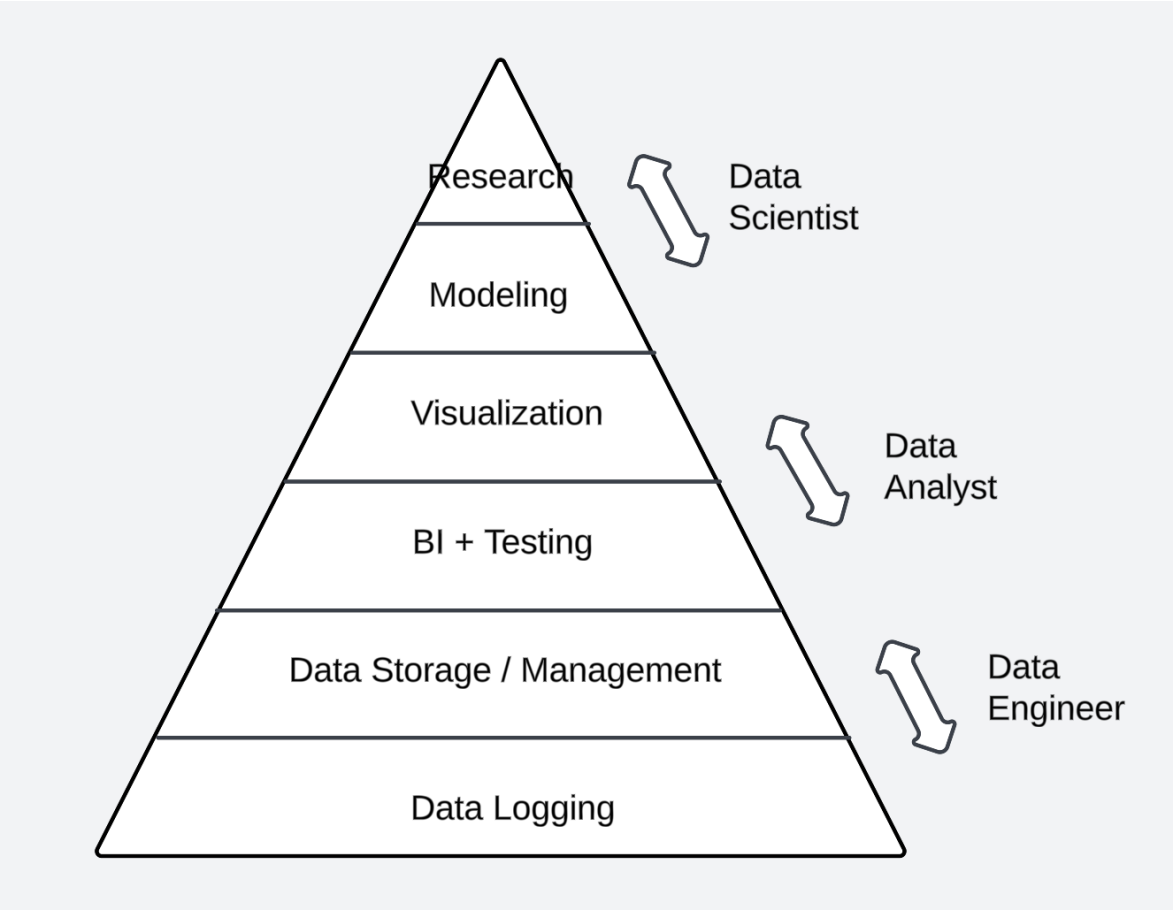
N – Interpret results

Data science can be used to make diagrams, graphs, and charts to represent trends and predictions. Data summarization helps stakeholders to implement results effectively” (Amazon Q, n.d., https://lc.cx/nDijJG).

2.1.2 Intro to three pillars of our discussion

For our discussion, we will focus on the three main pillars of data science in sports organizations and how they overlap within organizations. How can teams employ employees for these roles, and what education and background experience are required to get these roles?

Figure 1. Leading roles of data science



Source: own elaboration.

Data science-related opportunities at clubs fall into three leading roles: data engineer, data analyst, and data scientist. Each role serves a unique purpose. Put together, they form the backbone of any robust data science team. Understanding these roles and their responsibilities can help sports teams make strategic decisions about which type of expertise they need in their organization.

Data engineers are responsible for building the infrastructure and architecture for data organization. They create data pipelines that transform raw data into an easily understood format. In sports, data engineers ensure that the vast amounts of data collected from games, player statistics, and training sessions are efficiently stored for data analysts and data scientists. They use tools like Apache Hadoop and Apache Spark and databases like SQL and NoSQL to manage and streamline data flow.

Data analysts are storytellers whose primary role is interpreting data, analyzing results, and providing actionable insights. They work with raw and structured data sets, using statistical tools and data visualization platforms like Tableau and Power BI. For example,

analysts might examine player performance data in a sports team to identify strengths and weaknesses or analyze fan engagement metrics to improve marketing strategies.

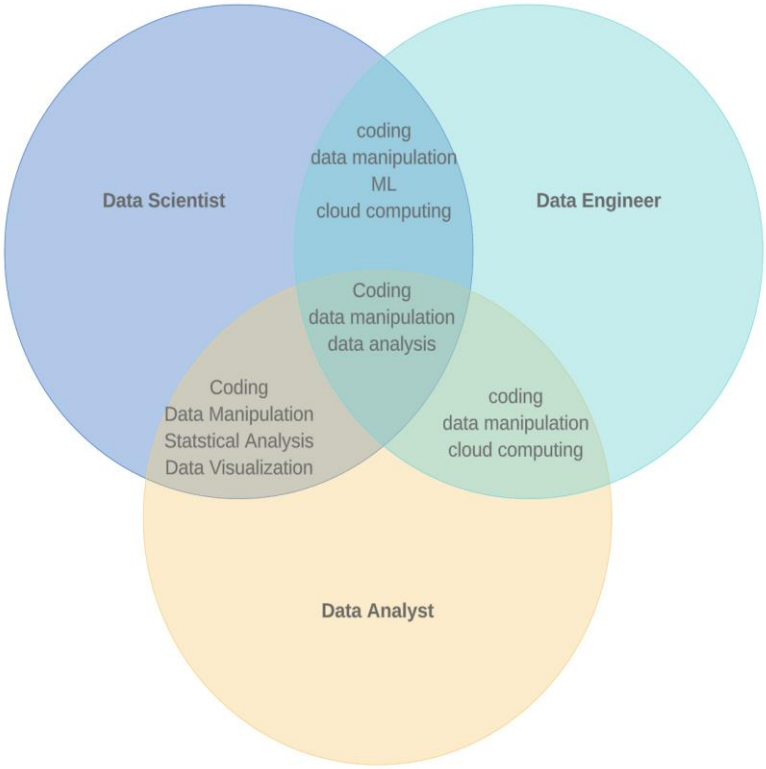
Data scientists are the connective tissue between data analysts and data engineers. They process and analyze data to build predictive models and machine learning algorithms to forecast outcomes. In sports, data scientists might develop models to predict the likelihood of injury, simulate game scenarios, and forecast player potential. They rely heavily on programming languages like Python and R and machine learning frameworks like TensorFlow and sci-kit-learn.

Table 1. Skills required

Data analyst	Data Engineer	Data scientist
Data visualization	Data warehousing & ETL	Data mining (EDA)
Adobe & Google Analytics	Data architecture & pipelining	Statistical & Analytical skills
Scripting & Statistical skills	Advanced programming knowledge	In-depth programming knowledge (SAS/R/ Python coding)
Reporting & data visualization	In-depth knowledge of SQL/ database	Hadoop-based analytics
Programming knowledge	Hadoop-based analytics	Machine Learning & deep learning principles
SQL/ database knowledge	Machine learning concept knowledge	Data optimization

Source: own elaboration.

Figure 2. Similarities and differences in skillset



Source: own elaboration.

Unit 2.2 Introduction to Data Engineering

2.2.1 Definition

Data engineering is a critical function within the data science and analytics field. Data engineers are responsible for designing, constructing, and maintaining the infrastructure for the organizations that allows them to collect, store, and process large amounts of data. Their work ensures data is accessible, organized, and ready for analysis by analysts and scientists.

2.2.2 Education and skills required for data engineering

Educational background

Most data engineers have a solid educational background in computer science and information technology. A bachelor's degree is typically the minimum requirement, but

many data engineers also hold a master's degree in computer science or software engineering.

Technical skills

Programming languages: data engineers must be proficient in Python and SQL. Knowledge of database systems like MySQL, PostgreSQL, Oracle, and NoSQL databases such as MongoDB and Cassandra is essential. Experience in cloud platforms like Amazon AWS, Microsoft Azure, Google Cloud, etc., is needed as well. Understanding how to design, implement, and manage databases is a core skill for data engineers. They need to be able to build data ingestion and transformation pipelines to read raw data from data provider APIs and transform the data into a structured database schema. For dealing with big data, where the datasets are 100s of Giga Bytes or more, data engineers need knowledge and experience using technologies like Hadoop and Spark.

2.2.3 The role of data engineers in the data ecosystem

Data engineers work closely with data scientists, analysts, and other IT professionals to ensure that data is available in a format ready for analysis. They also play a vital role in setting up the systems and processes that allow for efficient data handling, from raw data ingestion to the creation of data warehouses and data lakes.

Importance of data engineering

Organizations rely heavily on data to make informed decisions in today's data-driven world. Data engineers work to organize, make data consistent, and make it easier to access. They create pipelines that allow data to flow smoothly from various sources into a central repository for analysis.

Example: a data engineer's role is to collect and organize data into the data cloud or data lake used by the teams, where the data analyst can connect and use the columns to answer the stakeholders' questions.

2.2.4 Day-to-day duties of a data engineer

Building and maintaining data pipelines

The data engineer's primary responsibility is creating and maintaining data pipelines. These pipelines extract data, transform it into a usable format, and store it in a data warehouse. This process is called ETL (extract, transform, load).

Example: Apache Hadoop is an open-source framework for processing and storing large datasets in a distributed computing environment. Apache Spark is a fast in-memory data processing engine that supports SQL, streaming, machine learning, and graph processing.

Data warehousing

Data engineers are often responsible for designing and managing data warehouses, which are large-scale databases optimized for querying and analysis. This involves organizing data for easy access and sharing.

Example: Amazon Redshift, a fully managed data warehouse service in the cloud optimized for running complex queries on large datasets. Google BigQuery, is a serverless, highly scalable, cost-effective multi-cloud data warehouse designed for business agility.

Cloud platforms

Amazon Web Services (AWS): offers a suite of cloud services, including storage, computing, and database management, which data engineers widely use.

Google Cloud Platform (GCP): provides various cloud-based tools for data processing, storage, and machine learning.

Ensuring data quality and integrity

Data engineers must ensure the data being processed is accurate, consistent, and error-free. This involves setting up validation checks, cleansing data, and implementing best practices for data management.

Collaborating with data scientists and analysts

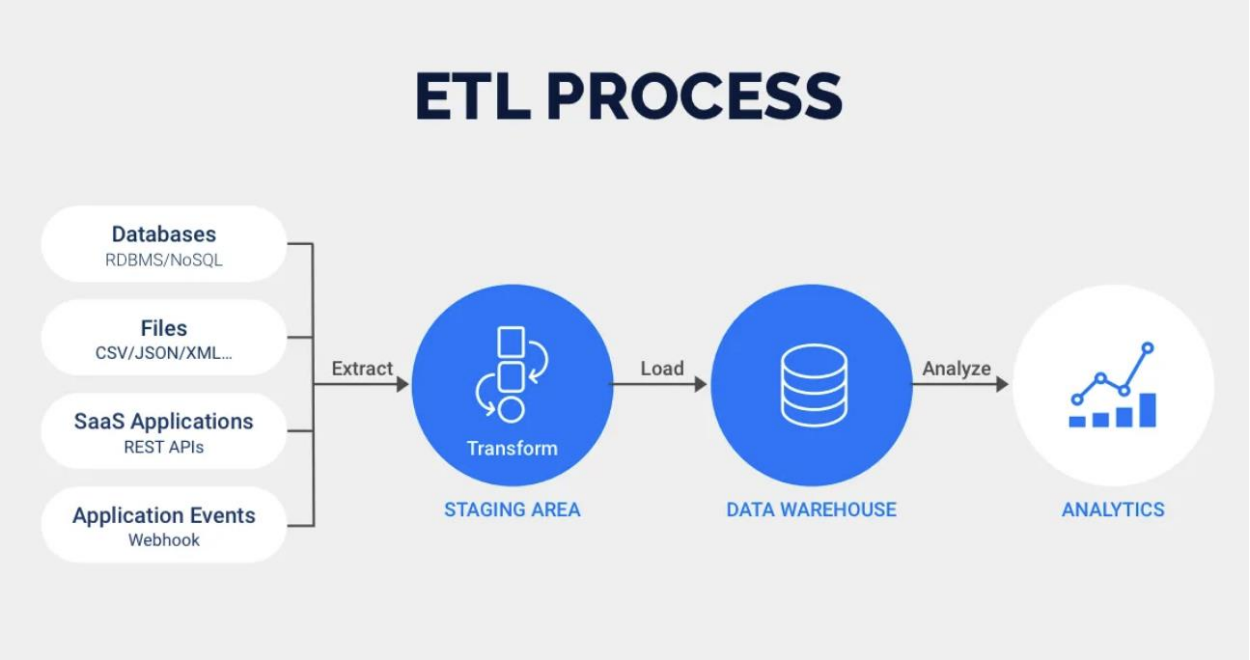
Data engineers often work closely with data scientists and analysts to understand their data needs. They provide the infrastructure for these professionals to perform their analyses efficiently and effectively.

Example: data visualization and BI tools. Tableau is a popular data visualization tool that allows users to create interactive and shareable dashboards. Power BI, a business analytics tool by Microsoft that provides interactive visualizations and business intelligence capabilities.

Monitoring and optimizing data systems

Data engineers are responsible for monitoring the performance of data systems and making necessary adjustments to optimize their efficiency. This might involve scaling systems to handle increased data loads or optimizing queries to run faster.

Figure 3. ETL process



Source: Surve, 2023, https://lc.cx/p6b1qv

2.2.5 Case studies: data engineers in sports

Take Statsbomb, for example. They provide detailed technical data about players and their performance during games. Teams can access this data through Statsbomb's website. But they also offer access to their API, how teams use this data is entirely up to them. A data engineer might download all the Statsbomb data and store it in the team's chosen services, like the Cloud or an on-prem database. This setup means teams are independent of Statsbomb for analysis. A data engineer creates a comprehensive dataset by combining physical data from GPS with technical data. Analysts can use this final product to assess players' physical capabilities and technical game efficiency. While doing this for one game in Excel is manageable, imagine handling data for ten-plus leagues with all their players. This approach offers an in-depth analysis of a player's performance over time, revealing valuable insights. This is the value that a data engineer can add to a team, as you will have multiple streams of data flowing in, and having access to all the data can help you make better decisions or at least show more possibilities.

2.2.6 Variations in roles across companies and industries

Data engineering in different industries

The role of a data engineer can vary significantly depending on the industry. In finance, data engineers may focus on ensuring the security and accuracy of financial data. In contrast, in e-commerce, they focus on building systems to handle large volumes of customer data and transactions. Data engineers might have broader responsibilities in smaller companies or startups, covering data architecture and analytics. The role will

always be more specialized in larger organizations, with different engineers focusing on specific aspects of data infrastructure, such as data warehousing, pipeline development, or cloud data solutions.

The evolving role of data engineers

As companies increasingly adopt cloud computing and big data technologies, the role of data engineers continues to evolve. There is a growing demand for engineers skilled in cloud-based data solutions and big data processing frameworks. Additionally, the rise of machine learning has led to new opportunities for data engineers to work closely with data scientists in building and deploying models. Data engineering is a dynamic and essential field that enables organizations to leverage data for decision-making and innovation. The role of a data engineer is multifaceted, requiring a blend of problem-solving skills and the ability to adapt to the changing demands of the industry.

Unit 2.3 Introduction to Data Analyst

2.3.1 Definition

A data analyst collects, processes, and performs statistical analyses on small and large datasets to help organizations make informed decisions. They are crucial in translating numbers and data into actionable insights that can influence business strategies, product development, and marketing campaigns. They are the storytellers of the data ecosystem.

2.3.2 Education and skills required for data analyst

Educational background: most data analysts have a background in mathematics, statistics, computer science, economics, or a related field. A bachelor's degree is typically required, but many data analysts also pursue advanced degrees in data science, business analytics, or applied statistics to deepen their expertise.

Technical skills: SQL is essential for querying databases and retrieving data. Some level of coding in a language like Python will be convenient but not required.

Statistical analysis: it is crucial to have a strong understanding of statistical concepts and techniques, including hypothesis testing, regression analysis, and probability distributions.

Data visualization: skills in data visualization tools such as Tableau, Power BI, and Excel are essential for creating compelling visual representations of data.

Beyond the above skills, data analysts are usually subject matter experts. In a sport like football, data analysts need to understand the sport, tactics, strategies, etc., employed in football.

2.3.3 The role of data analysts in the data ecosystem

Data analysts work across various departments, including finance, marketing, operations, and product development. They bridge raw data and strategic decision-making, ensuring data-driven insights are at the forefront of organizational strategies. Their role often involves collaborating with data scientists, engineers, and business stakeholders. In a world where data is generated at an unprecedented rate, the ability to analyze and interpret data has become essential for businesses to stay competitive. Data analysts help organizations understand trends, patterns, and correlations within data, allowing them to predict future outcomes, optimize operations, fill gaps between the process and outcomes, and identify new opportunities.

2.3.4 Day-to-day duties of a data analyst

Data collection and cleaning

Data analysts gather data from various sources, including databases, spreadsheets, and external data feeds. One of the first tasks is to clean the data, removing inconsistencies, errors, and duplicates to ensure the dataset's quality.

Example: data analysis tools. Python, a versatile programming language widely used for data manipulation, statistical analysis, and machine learning. Libraries like Pandas, NumPy, and SciPy are essential for data analysis. R, another programming language prevalent in statistics and data analysis, known for its extensive range of statistical and graphical techniques. Microsoft Excel, Tableau and Power BI are other tools that data analysts can use to analyze data without writing code.

Data exploration and analysis

Once the data is clean, analysts explore it to understand its structure and identify patterns or trends. This exploration phase often involves descriptive statistics, visualizations, and summary statistics to gain insights into the data's distribution and characteristics.

Example: data querying tools. Structured Query Language (SQL) is the standard language for querying and managing databases. Data analysts use SQL to retrieve, filter, and aggregate data from relational databases like MySQL and Cloud storage.

Data visualization

Data analysts create visual representations of data to help stakeholders understand complex information more easily. They use charts, graphs, and dashboards to convey critical findings, trends, and correlations. Tools like Tableau, Power BI, and Excel are mainly used.

Example: data visualization tools. Tableau, is a leading data visualization tool that allows users to create interactive dashboards and reports. It also provides powerful visualization capabilities. Power BI, a business analytics tool by Microsoft enables users to visualize data and share insights across an organization. It integrates seamlessly with other Microsoft products. Excel, despite being an older tool, remains a data analysis and visualization staple. Its versatility and familiarity make it a go-to tool for many analysts.

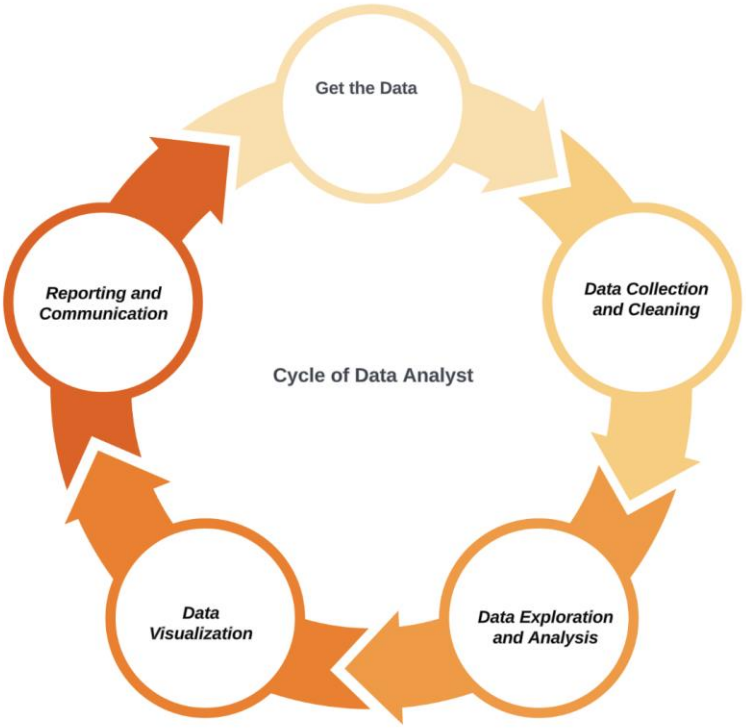
Reporting and communication

A significant part of a data analyst's job is communicating findings to non-technical stakeholders. This involves creating reports and presentations summarizing the analysis and providing clear and actionable insights. Effective communication is critical to ensuring that the insights are understood and implemented.

Collaborating with other teams

Data analysts often work closely with other teams, such as marketing, sales, and product development, to understand their data needs and provide tailored analysis. They also collaborate with data engineers to ensure the data infrastructure supports their analytical requirements. Data analysts may be responsible for tracking key performance indicators (KPIs) and metrics over time. They monitor the effectiveness of strategies and campaigns, providing feedback and suggestions for improvement based on the data.

Figure 4: Cycle of data analyst



Source: own elaboration.

2.2.5 Case studies: Data Analyst in Sports

Take player health data, for example. A data analyst dives deep into player health data, turning numbers into meaningful insights. By analyzing loads and tracking trends over weeks, they can spot subtle changes in performance and recovery. A slight dip in speed or a minor increase in heart rate might seem insignificant in isolation. Still, when visualized over time, these details can reveal patterns, like signs of overtraining or a developing injury. Data analysts must be good at mathematically proving their results and have a keen eye for storytelling. They will combine those two skills and build an impactful dashboard or chart to convey a message to the stakeholders. These insights help coaches make informed decisions to keep players at peak performance. Imagine having multiple streams of data and a storyteller who understands the data and answers all your questions by analysing the data finding the KPI's with a visual to show the impact of the data.

2.3.6 Variations in roles across companies and industries

Data analysis in different industries

The role of a data analyst can vary significantly depending on the industry. Data analysts may focus on analyzing market trends, risk management, and financial modeling in finance. They might analyze customer behavior, campaign performance, and sales data in marketing to optimize marketing strategies. They could also work on patient data,

clinical trial results, and healthcare outcomes to improve patient care and operational efficiency. As data's importance continues to grow across industries, the demand for skilled data analysts is rising. Organizations increasingly recognize the value of data-driven decision-making and invest in solid data analytics teams. The role of a data analyst is expanding beyond traditional data analysis techniques to include skills in machine learning, big data technologies, and cloud computing.

The evolving role of data analysts

As companies increasingly rely on data to drive their strategies, the role of data analysts continues to evolve. There is a growing demand for analysts who are not only skilled in traditional data analysis techniques, but also in advanced analytics, machine learning, and data engineering. The rise of big data and cloud computing has also led to new opportunities and challenges for data analysts. Data analysts face several challenges, including working with large, complex datasets, integrating data from various sources, and providing real-time insights. Future trends in data analysis include the increasing use of automation and AI to enhance analysis capabilities, the growing importance of data ethics and privacy, and the continued expansion of data-driven decision-making across all sectors. Data analytics is becoming a core component of business strategy, influencing everything from product development and marketing to operations and customer service. Data analysts play a crucial role in this transformation by providing the insights that drive strategic decisions.

Unit 2.4 Introduction to Data Scientist

2.4.1 Definition

Data science is a multidisciplinary approach combining statistics, mathematics, and computer science knowledge with domain knowledge to extract meaningful insights from data. A data scientist is a professional who uses data to solve complex problems, builds predictive models, and drives decision-making.

2.4.2 Education and skills required for data science

Educational background

Data scientists typically have an educational background in computer science, statistics, mathematics, or engineering. A master's degree is often the minimum requirement, but many data scientists also hold advanced degrees (PhD) in data science or statistics.

Technical skills

Programming languages: data scientists must be proficient in programming languages like Python and R. These languages are widely used for data manipulation, statistical analysis, and machine learning.

Statistical and analytical skills

Statistical analysis: data scientists need a solid foundation in statistics. They must be expert with probability distributions, hypothesis testing, and inferential statistics, time series analysis, multivariate statistics.

Analytical thinking: data scientists need strong analytical skills to identify patterns, correlations, and causations within datasets. Data scientists must possess a problem-solving mind to solve complex problems using data-driven approaches.

2.4.3 The role of data scientists in the data ecosystem

In today's digital age, organizations generate vast amounts of data from various sources, including customer interactions, social media, sensors, and transactions. Data science allows organizations to harness this data to gain a competitive advantage, optimize operations, and innovate. Data scientists are crucial in transforming raw data into valuable insights that drive business growth.

Data scientists usually process and analyze large datasets to uncover trends, patterns, and insights that can inform positive business decisions. They develop algorithms, build predictive models, and use advanced statistical techniques to solve complex problems. Data scientists often work closely with data engineers, analysts, and business stakeholders to ensure their findings are actionable and aligned with organizational goals.

2.4.4 Day-to-day duties of a data scientist

Data collection and preparation

They are responsible for cleaning and preprocessing the data to ensure it is ready for analysis. This involves handling missing values, normalizing data, and transforming it into a format suitable for modeling.

Exploratory data analysis (EDA)

Exploratory data analysis is a critical step in understanding the underlying patterns and relationships in the data. Data scientists use statistical techniques and visualization tools to explore the data, identify trends, and detect anomalies. EDA helps data scientists form hypotheses and guides the development of predictive models.

Example: data manipulation tools. Python is the most popular programming language for data science. It is known for its simplicity and a wide range of libraries, such as Pandas, NumPy, and Scikit-learn. R is another widely used language in data science, particularly for statistical analysis and visualization. It offers various packages, like ggplot2, for creating complex visualizations.

Model development and machine learning

A significant part of a data scientist's role is building predictive models using machine learning algorithms. Based on the problem, they select appropriate models, such as linear regression, decision trees, or neural networks. Data scientists train these models on historical data and evaluate their performance.

Example: machine learning frameworks. Scikit-learn, is a powerful Python library for machine learning, offering tools for data mining, data analysis, and building predictive models. TensorFlow developed by Google, is a popular open-source framework for deep learning. Keras, built on top of TensorFlow, simplifies the process of building and training deep learning models.

PyTorch, developed by Facebook, is an open-source machine learning framework prevalent for deep learning research and development.

Model evaluation and optimization

Once a model is built, data scientists evaluate its performance and fine-tune it to improve accuracy. This may involve hyperparameter tuning, cross-validation, and feature engineering. The goal is to ensure the model generalizes well to new, unseen data.

Communication and reporting

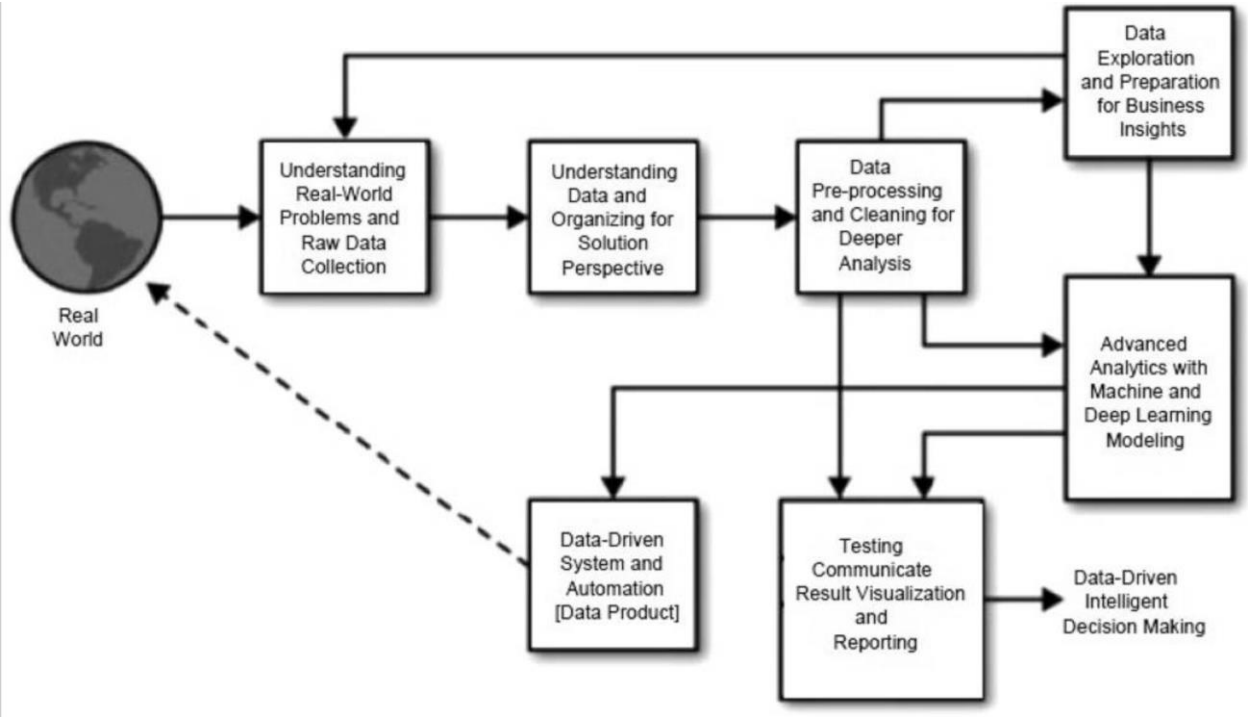
Data scientists must effectively communicate their findings to non-technical stakeholders. This involves creating visualizations, dashboards, and reports that convey the insights derived from the data. Data scientists often present their results to senior management, explaining the implications and potential actions based on their analyses.

Example: data visualization tools. Matplotlib and Seaborn, are two Python libraries that are commonly used to create static, animated, and interactive visualizations in Python. Tableau, is a leading data visualization tool that allows data scientists to create interactive and shareable dashboards. Tableau is particularly useful for presenting data-driven insights to business stakeholders.

Collaboration with other teams

Data scientists frequently collaborate with other teams, including data engineers, analysts, and business leaders. They align data science projects with business objectives and ensure the data infrastructure supports their analytical needs.

Figure 5. An example of data science modeling from real-world data to data-driven system and decision-making



Source: Sarker, 2021, https://lc.cx/8_g88J

2.2.5 Case studies: Data Scientist in Sports

Data scientists collect and analyze data from various sources, including player tracking systems, wearables, and video analysis. Moreover, data scientists help coaches understand player movements, shot selection, and defensive patterns in team sports like soccer and basketball. These insights enable teams to develop more effective game plans and make real-time decisions during matches. Data scientists also play a crucial role in recruitment by analyzing player statistics and performance metrics to identify talent that fits the team's needs. Data science enhances fan experiences for the front office by personalizing content, improving ticket sales strategies, and optimizing stadium experiences. Overall, the role of a data scientist in sports is multifaceted, focusing on optimizing performance, ensuring player health, and enhancing the overall experience for fans and stakeholders.

2.4.6 Variations in roles across companies and industries

Data science in different industries

The role of a data scientist can vary significantly depending on the industry. In finance, data scientists may focus on developing predictive models for risk management and algorithmic trading. In healthcare, they might analyze patient data to improve outcomes and develop personalized treatment plans. Data scientists often analyze customer behavior in retail to optimize pricing strategies and enhance the customer experience. In startups or smaller companies, data scientists take on a broader role, handling everything from data collection and cleaning to model development and deployment. They may also be involved in business strategy and decision-making. The role might be more specialized in larger organizations, with different teams focusing on specific aspects of data science, such as data engineering, machine learning, or data visualization.

The evolving role of data scientists

As data science evolves, data scientists are increasingly expected to have a broad skill set, including data analysis, machine learning, data engineering, cloud computing, and AI. The demand for data scientists with expertise in deep learning, natural language processing, and big data technologies is growing as organizations seek to leverage more complex and sophisticated data. Data scientists face several challenges, including the need to handle large, complex datasets, the integration of disparate data sources, and the need for real-time analytics. Future trends in data science include the increasing use of automation and AI to streamline the data science process, the growing importance of data ethics and privacy, and the continued expansion of data-driven decision-making across all sectors.

References

Amazon Q. (s.d.). What is Data Science? https://aws.amazon.com/what-is/data-science/#:~:text=Data%20science%20is%20the%20study,analyze%20large%20amounts%20of%20data.

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Comput. Sci*, 2(5), 377. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8274472/

Surve. S. (2023, Febraury 27). 30 days of Data Engineering: Day 1. *Medium*. https://sarangsurve.medium.com/30-days-of-data-engineering-day-1-290795b0da85

Bibliography

Martin, R. C. (2018). *Clean Code: A Handbook of Agile Software Craftsmanship*. Pearson.

By Robert C. Martin is an excellent book for understanding the principles of writing good, maintainable code.

Sweigart, A. (2015). *Automate the Boring Stuff with Python.* No Starch Press.

By Al Sweigart is a great introductory book that makes learning Python fun and practical.

Williams, H. (2018, January 23). The Pyramid of Data Needs (and why it matters for your career). *Medium*. https://medium.com/@hugh_data_science/the-pyramid-of-data-needs-and-why-it-matters-for-your-career-b0f695c13f11