



Paper PD07

Bridging the Gap: Role Transition from SAS Programmer to Data Scientist

Shashikant Kumar, Efficacy Lifescience Analytics, Bengaluru, India
Mrityunjay Kumar, Efficacy Lifescience Analytics, Bengaluru, India

Abstract

Often, we hear people talking about Data Science and we always get overwhelmed with multiple open-ended questions such as what it is, who can do it, what needs to be done? Can SAS Programmers become Data Scientist? Will it be a wistful decision? what else to expect in role transition etc. We have also witnessed that in the last few years, demand for Data Scientists have increased significantly, and how clinical healthcare industry remains untouched by presence of Data Scientist. Moving from a primarily procedural and structured environment to one driven by machine learning, big data, and advanced analytics entails a significant shift in mindset and skill set. This paper outlines the essential components of Data Scientist skills and usage of open-source software that can be leveraged along with challenges, opportunities, real-world applications, elaborate on logical differences based on real time examples. This may benefit the industry in coming years.

Introduction

The role of data scientists has become a cornerstone in the digital era, with organizations leveraging advanced analytics to drive decision-making. The skillset of a SAS programmer, particularly in the pharmaceutical industry, includes in-depth statistical knowledge, familiarity with clinical trial data, and a structured approach to programming. However, the role of a data scientist expands these requirements to include proficiency in machine learning, data visualization, and working with big data platforms which may include open-source technologies as well.

Data science is an interdisciplinary field that extracts knowledge and insights from many structured and unstructured data, using scientific methods, data mining techniques, machine-learning algorithms, and big data. It plays a crucial role in clinical trials to support the collection, management, and analysis of clinical data. The adoption of data science in clinical trials is seen across the globe. It enables the use of predictive modelling and machine learning to identify suitable patient populations, forecast dropout rates, and streamline trial workflows. Diverse data sources like electronic health records (EHRs), wearables, and real-world evidence, may lead to more accurate and timely insights. Additionally, Data science facilitates adaptive trial designs and risk-based monitoring, improving patient safety and thereby reducing costs. By accelerating decision-making and enabling the discovery of personalized treatment strategies, Data science significantly enhances the efficiency, quality, and success of clinical trials, contributing to better healthcare outcomes.

Understanding the Skills Gap

As digital transformation accelerates, businesses must adopt a skills-based approach to learning, leveraging AI and tech platforms to identify and bridge skills gaps, thereby boosting employee retention and business performance.

Businesses are changing faster than ever, making continuous learning not merely a nice perk but an absolute necessity. With the advent of digital transformation and the increasing reliance on technology, employees must continually update their skills to stay relevant.

The skills gap is a multifaceted issue, encompassing a lack of technical expertise, soft skills, and industry-specific knowledge.



Source: analyticsweek.com

Some of the key factors contributing to this gap include:

1. **Technological Advancements:** Rapid technological innovation creates a demand for new skills that the current workforce may not possess. According to the World Economic Forum, half of the global workforce will likely need reskilling by 2025. As experienced employees retire, there is a gap left in their wake that is hard to fill. Bain & Company projects that, by 2031, workers aged fifty-five and older will exceed a quarter of the workforce.
2. **Skill Gaps:** The education system sometimes lags industry needs, producing graduates without the necessary skills for today's job market. Roughly half of university graduates end up in jobs that do not make use of their skills.
3. **Changing Job Roles:** Jobs are evolving, and roles that did not exist a few years ago are now critical, requiring a different skill set. The World Economic Forum predicts that eighty-five million jobs may be displaced by a shift in the division of labor between humans and machines, while ninety-seven million new roles may emerge that are more adapted to the new division of labor.

Understanding the Role Transition from SAS Programmers to Data Scientist

This paper explains various focus areas which can be considered for the role transition based on current trends and future requirements in Pharmaceutical Industries as data science applications are at the forefront of many innovations.

Focus Areas	Challenges Faced in the Clinical Industry	Key Differences Between SAS Programming and Data Science Roles	Importance of Adaptability and Continuous Learning
Competency	Learning new programming languages (e.g., Python, R) and mastering machine learning frameworks.	SAS programmers focus on structured data and predefined protocols, while data scientists work with diverse data types and exploratory analysis.	Acquiring new skills in cloud computing, big data technologies, and machine learning is essential.
Data Handling	Limited experience with unstructured/semi-structured data (e.g., text, images) compared to structured clinical data.	SAS primarily handles structured clinical trial data, whereas data scientists often work with unstructured and complex data.	Understanding different data formats and applying advanced techniques like NLP for unstructured data.
Tools & Technologies	Adapting to new tools like TensorFlow, scikit-learn, Hadoop, and cloud-based platforms (AWS, Azure)	SAS is used for statistical programming, while data scientists use a variety of tools (Python, R, Spark) for predictive analytics and data engineering	Continuous learning about new technologies and tools is required to stay current and efficient in data science

Analytical Methods	Gaining proficiency in advanced analytics, machine learning, and deep learning, which require a shift from traditional statistical methods	SAS programming relies on traditional statistical models (e.g., PROC procedures), whereas data scientists employ advanced machine learning algorithms and AI	Learning new methodologies and techniques for building and evaluating models, such as neural networks and decision trees
Project Scope	Difficulty in understanding the broader scope of end-to-end data science projects, including data engineering and deployment	SAS programmers often focused on regulatory submissions, whereas data scientists design and deploy predictive models and solutions end-to-end	Developing expertise beyond statistical programming to include data engineering, software development, and model deployment
Organizational Support	Resistance to role transition due to lack of mentorship, guidance, or defined career paths	SAS roles are well-defined within clinical research, while data science roles are more flexible and interdisciplinary	Seeking mentorship and networking opportunities to gain guidance and support during the transition
Continuous Learning	Keeping up with evolving technologies, methodologies, and tools in data science can be overwhelming without structured learning paths	SAS roles often have a structured scope, while data scientists must continuously explore and learn new approaches and tools	Attending workshops, participating in various conferences are critical for growth

Adapting Skills for Data Science

1. Tools and Technologies

For years, Microsoft Excel and SAS have been the preferred data analysis tools for professionals in every setting, including clinical research and trials during drug development. While they both have been effective in past research efforts, the acceleration of data collection and processing power has led to a revolution of new flexible tools.

As illustrated above, the role transition from a SAS programmer to a Data Scientist requires expanded skills that goes beyond traditional SAS programming knowledge. SAS programmers are already familiar with data manipulation, statistical analysis, and reporting. However, to excel in a Data Scientist role, additional skills in programming, data engineering, and advanced analytics are essential. Below are the key technical skills required for this role transition:

- 1. Machine Learning and AI Techniques:** SAS programmers focus on traditional statistical methods and data manipulation but may lack exposure to machine learning algorithms, such as decision trees, random forests, and deep learning techniques, which are essential for predictive modelling in data science.
- 2. Big Data Technologies:** Unlike SAS environments, which often handle structured data, data scientists frequently work with unstructured data and big data frameworks (like Hadoop and Spark) to analyse large-scale datasets. SAS programmers may not be accustomed to these tools.
- 3. Leveraging Open Source:** to take advantage of big data analytics, organizations have turned to alternatives such as R programming and Python which offers advanced data analysis functions and capabilities. R programming can handle larger data sets and generate more detailed visualizations. In Python, data scientists can manipulate data like functions in Excel and import data like SAS. However, they can also make use of advanced statistics and machine learning capabilities. Python can scale and work with more extensive and multiple datasets. It can also handle large volumes of data at faster processing speeds, which is suited well for data storage of thousands of patient entries and millions of tables.
- 4. Cloud Computing:** Data Scientists increasingly utilize cloud platforms (AWS, Azure, Google Cloud) for scalable data storage, processing, and machine learning model deployment. SAS programmers may not have experience with cloud environments and services.
- 5. Version Control and Collaborative Tools:** Familiarity with version control systems (like Git) is crucial for collaborative projects in data science. SAS programmers might not regularly use such tools for code management and collaboration, limiting their ability to work in diverse teams.

6. **Automation of Workflow:** Automating data workflows using tools like Apache Airflow or writing scripts in Python for ETL processes is vital in data science but may not be part of a SAS programmer's skillset.
7. **Domain-Specific Knowledge in Advanced Analytics:** Understanding advanced analytics concepts like Natural Language Processing (NLP) or image analysis for clinical data (e.g., medical imaging) is important for data scientists, while SAS programmers may not typically cover these areas.

2. Soft Skills

Soft skills are crucial for effectively communicating data-driven insights, collaborating with multidisciplinary teams, and driving project success. The following soft skills are key for Data Scientists, especially in clinical research and healthcare fields:

1. **Effective Communication and Collaboration:** Clinical Stakeholders such as Clinicians, Statisticians, Regulatory Agencies, and others may have different needs and expectations which involves an effective collaboration. Communicate findings and insights in a way that supports clinical decisions and patient care. Example: Survival analysis results - explain about median survival time and hazard ratios for different patient subgroups.
2. **Adaptability and Learning Agility:** One should be adaptable and willing to continuously learn new methods, tools, and industry trends and changing needs of regulatory landscapes. Embrace new technologies like Machine Learning (ML) and Artificial Intelligence (AI) as they become relevant to clinical data analysis.
3. **Domain Knowledge: Understanding Clinical Trial Protocols and Regulatory Requirements**
Clinical trial protocols define the objectives, design, methodology, statistical considerations, and organization of a clinical trial. They outline the rationale behind the study, participant criteria, treatment schedules, and endpoints. Ensure that any deviations from the protocol are handled according to predefined criteria and documented accordingly. Familiarity with these elements is essential for Data Scientists to properly analyse and interpret data.

Regulatory agencies like the FDA, EMA, PMDA etc. provide guidelines for conducting clinical trials, ensuring participants safety, and reporting results. Data Scientists must be familiar with these guidelines, as well as industry standards like CDISC, which define how clinical trial data should be structured and submitted. SDTM (Study Data Tabulation Model) and ADaM (Analysis Data Model) are created in accordance with CDISC standards, facilitating seamless regulatory review and submission.

Relevance of SAS Programming Skills in Data Science

SAS programming skills serve as a strong foundation for various data science activities. Below table demonstrates the transferable skills and applications of SAS programming in data science projects that can be leveraged.

Leveraging SAS Skills	Description	Application in Data Science
Data Management and Manipulation	SAS programmers excel in handling large datasets and applying complex data transformations.	Data wrangling and preprocessing are essential steps in building machine learning models.
Statistical Analysis and Model Building	SAS procedures like PROC REG, PROC LOGISTIC, and PROC MIXED offer a strong foundation in statistical techniques.	These skills help in transitioning to machine learning models such as regression, decision trees, and clustering.
Handling Complex Data Structures and Missing Data	Expertise in dealing with hierarchical data, repeated measures, and missing data imputation.	Applicable in hierarchical clustering, time-series forecasting, and predictive modeling.
Automation and Reporting	Proficiency in automating workflows and generating reports using macros and the ODS system.	Automation of data science workflows, reproducible reporting using Jupyter Notebook or RMarkdown.
Compliance and Data Privacy	Familiarity with industry standards like CDISC, HIPAA, and GDPR for clinical data.	Ensures data science projects adhere to compliance and privacy guidelines.

Some Examples of Real-world Applications as Data Scientists

Data science has played a transformative role in the clinical research industry, enhancing drug development, patient care, and regulatory submissions. From accelerating clinical trials to identifying new drug targets and predicting patient outcomes, data science projects have contributed significantly to various aspects of healthcare and research. Here are some notable examples of successful data science projects in clinical research:

Predictive Analytics for Patient Recruitment and Retention

One of the critical challenges in clinical trials is patient recruitment and retention. One of large pharmaceutical companies utilized predictive analytics to improve patient recruitment strategies for their trials. By analysing historical clinical trial data and patient demographics, they developed models to predict the likelihood of patient enrolment at specific trial sites. This project significantly reduced recruitment time and costs. It enabled them to target the right patient populations and improve overall trial efficiency, ensuring that studies were completed on schedule and within budget.

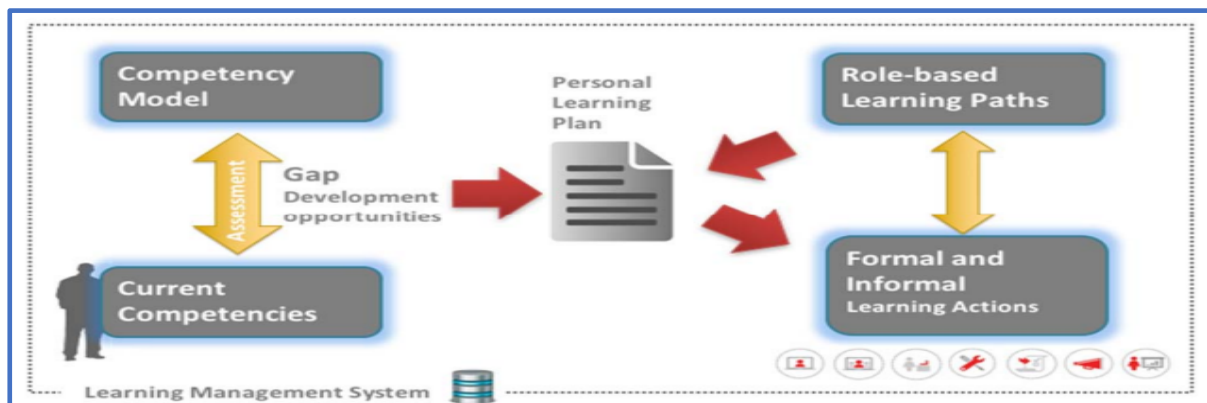
Some other examples:

- Machine Learning Models for Predicting Disease Progression
- Real-World Data (RWD) Analysis for Drug Repurposing
- AI-Driven Clinical Trial Monitoring and Risk-Based Monitoring (RBM)
- Natural Language Processing (NLP) for Adverse Event (AE) Detection
- Image Analysis Using Deep Learning in Oncology Trials
- Digital Biomarkers for Neurological Disorders
- AI-Powered Drug Discovery and Development

Future Opportunities

1. Professional Development Cycle

As SAS programmers transition into Data Science roles, a structured development path is essential. Career progression involves gap analysis, learning new technologies and adapting to broader data science practices while leveraging the strong analytical skills SAS programmers already possess.



Source: SAS.com

Competency-Based Learning

2. Stages of Professional Growth:

1. Entry Level (SAS Programmer/Data Analyst):

Focus: Expanding from SAS to tools like R, Python, and SQL.

Key Skills: Data manipulation, basic machine learning models, familiarity with clinical data structures (SDTM/ADaM).

2. Intermediate Level (Junior Data Scientist):

Focus: Learning advanced data science techniques, including supervised and unsupervised learning, deep learning, and predictive analytics.

Key Skills: Proficiency in statistical modeling (linear and logistic regression, survival analysis), data visualization, and familiarity with cloud computing (AWS, GCP).

3. Advanced Level (Senior Data Scientist):

Focus: Strategic thinking, leading data science projects in clinical trials, AI/ML applications for healthcare, and contributing to regulatory submissions using machine learning insights.

Key Skills: Leadership, communication with cross-functional teams (clinicians, statisticians), and applying ethical AI practices in healthcare.

3. Continuing Education and Certification Programs

Staying updated on the latest technologies and methodologies is critical for SAS programmers moving into data science roles. Continuing education provides a platform for acquiring new skills, while certifications validate these competencies.

- **Online Courses and Bootcamps:**
 - Data Science and Machine Learning: Introductory to advanced courses in Python, R, machine learning, and AI.
 - Healthcare Data Analytics Specialization: Focuses on applying data science in the healthcare field.
- **Certifications:**
 - Certified Data Scientist: Provides credentials for mastery in data science, covering machine learning, big data, and AI.
 - SAS Advanced Analytics Certification: Expands existing SAS expertise to advanced analytics, machine learning, and deep learning.
 - CDISC Data Science Professional: Certification specific to applying data science in clinical trials, dealing with SDTM/ADaM and regulatory submissions.
 - Degrees in Data Science, Biostatistics, or Health Informatics from various institutions help professionals build deep technical expertise.

4. Networking and Mentorship Opportunities in the Clinical Research Community

Building a network and seeking mentorship are invaluable to career advancement. Engaging with peers, joining professional organizations, and finding mentors helps SAS programmers' transition smoothly into data science roles.

- PHUSE: A platform for professionals in clinical data science to share knowledge, attend workshops, and present research.
- DIA (Drug Information Association): Provides global forums for regulatory science and data analytics discussions.
- PharmaSUG: Software Users Group of life science and health research professionals focused on the application of technological solutions in data analytics and regulatory support.

Challenges in Role Transition

1. **Steep Learning Curve:** Moving from a rule-based SAS programming environment to the dynamic landscape of data science can be daunting. The continuous evolution of machine learning frameworks necessitates a commitment to lifelong learning.
2. **Lack of Real-World Experience:** Many SAS programmers have extensive experience in structured data but limited exposure to unstructured data sources like text or images, which are common in the field of data science.
3. **Integration with Current Role:** Balancing existing responsibilities while upskilling can be challenging. Managing time effectively and prioritizing learning goals is crucial.

Conclusion

Data science is rapidly transforming the landscape of clinical trials by improving study design, enhancing patient recruitment and retention, enabling real-time monitoring, and optimizing data management thus increasing the need of data scientists. The adoption of data science methodologies not only accelerates the clinical trial processes but also enhances the reliability and quality of findings, leading to more effective and safer treatments for patients. As the pharmaceutical industry continues to evolve, data science will play an increasingly critical role in shaping the future of clinical research where SAS programmers can effectively use their clinical and programming knowledge by aforesaid means in role transition to Data Scientist.

The key to unlocking these powerful capabilities is to understand how to implement algorithms, data collection, storage, validation, and visualization to generate valuable insights. With professional training, organization can develop an expert data science team and level up existing employees to suggest data science projects and facilitate collaboration and communication.

Acknowledgments

Authors would like to extend their sincere thanks to Efficacy Lifescience Analytics for giving them an opportunity to write and present this paper. Any brand and product names are trademarks of their respective companies.

Recommended Reading

Kelleher, J. D., & Tierney, B. (2018). Data Science: A Comprehensive Overview. Springer.

Van Veen, M., "The Role of Communication Skills in Data Science," Journal of Data Science Practice, 2021.

Smith, J. and Brown, R., "Soft Skills for Data Scientists: Navigating the Healthcare Landscape," Clinical Data Science Review, 2022.

<https://www.hrdconnect.com/2024/03/05/bridging-the-skills-gap-the-role-of-upskilling-and-reskilling-in-modern-business/>

<https://analyticsweek.com/bridging-the-data-talent-gap-strategies-for-cultivating-a-future-ready-workforce/>

<https://www.leveluphcs.com/blog/8-strategies-to-bridge-the-skills-gap>

<https://support.sas.com/resources/papers/proceedings14/1486-2014.pdf>

<https://support.sas.com/resources/papers/proceedings17/0832-2017.pdf>

<https://support.sas.com/resources/papers/proceedings14/SAS286-2014.pdf>

[Data Science's Innovation in Pharmaceutical Clinical Trials \(datasociety.com\)](https://www.datasociety.com/)

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Shashikant Kumar, Director - Biostatistics & Programming

Efficacy Lifescience Analytics

Bengaluru, India

E-mail: shashikant.kumar@efficacy.com

www.efficacy.com

Mrityunjay Kumar, Associate Director - Biostatistics & Programming

Efficacy Lifescience Analytics

Bengaluru, India

E-mail: mrityunjay.kumar@efficacy.com

www.efficacy.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.