Data Scientist Knowledge and Skills Evaluation Towards a Data-Driven Research Methodology

Jing Lu
University of Winchester, UK

Jing.Lu@winchester.ac.uk

Abstract: The modern business world increasingly requires a higher level of data science expertise as well as abilities in problem solving and data analytics. Data science is a broad and fast-moving field of methods, processes, algorithms and systems to extract insights from data. The University of Winchester has followed the guidance of the Institute for Apprenticeships and Technical Education (IfATE) and worked in partnership with a number of international, national and regional employers in the design and development of its Data Scientist (integrated degree) programme, which leads to a Bachelor's degree in Data Science. This degree apprenticeship supports students in gaining the knowledge and skills that are in demand by employers today and into the future, where working in multi-disciplinary teams alongside domain experts will often be the norm. IfATE specifies an End-point Assessment (EPA) plan to enable the apprenticeship to be completed in accordance with its Data Scientist degree apprenticeship standard. This paper considers professional practice and competence in data science and links the processes used in completing the EPA with domain-based knowledge and expertise. It reviews representative solution methodologies before demonstrating the applicability of a data-driven research methodology to discover insights and achieve organisational goals.

Keywords: Data science degree apprenticeship, Professional practice, Knowledge and skills, Modelling and evaluation, Research methodology

1. Introduction

Data science is an interdisciplinary subject that integrates (e.g.) data analytics, computer science, software engineering, project management and research methodology to develop intelligence with respect to real-world problems (Cao 2017). The Institute for Apprenticeships and Technical Education (IfATE) in the UK works with employers to develop, approve, review and revise apprenticeships and technical qualifications – one such apprenticeship being the Data Scientist (integrated degree). In partnership with a number of international, national and regional employers, a Data Science degree apprenticeship was designed at the University of Winchester to support companies/organisations in attracting data scientists and retaining their top digital talent. This programme provides apprentices with the opportunity to develop their knowledge, understanding, technical skills and indeed confidence to operate successfully in a globalised work environment.

IfATE has specified the areas of Knowledge data scientists must understand, the Skills they are able to deliver and the Behaviours they should demonstrate (KSBs). On the other hand, the EU-funded EDISON (Education for Data Intensive Science to Open New science frontiers) project has focused on activities to establish the new profession of Data Scientist and developed an impressive collection of documents including their Data Science Competence Framework, Data Science Body of Knowledge and Data Science Model Curriculum (EDISON 2017a). This paper starts by emphasising the importance of research methodologies for the data scientist profession based on a review of EDISON Data Science competence groups. The paper moves forward by mapping KSBs defined by the IfATE Data Scientist (integrated degree) standard with the body of knowledge and competency groups from EDISON, aiming to provide an organised and structured way to evaluate the knowledge and skills required for a Data Science apprenticeship.

The final-year End-point Assessment (EPA) represents the ultimate capstone for the Data Science programme, taking place once the apprentice has met all the KSBs in the standard and passed all the taught modules. It includes a significant work-based project, which informs a corresponding report on planning and executing the project, and a portfolio of evidence from other real-work projects to inform a professional discussion with the independent assessor. This paper compares macro-solution methodologies for data science and analytics projects before presenting a data-driven research methodology linked to skills and behaviours specified by IfATE. The paper then concludes with some examples of data science applications in the workplace, highlighting the domain-based context of apprentices.

2. Data Scientist Professional Practice/Profiles

The modern business world increasingly requires a higher level of data science expertise as well as abilities in problem solving and data analytics. Data Scientists can have an impact at strategic and operational levels by building and maintaining collaborative relationships with key stakeholders, subject experts and colleagues as

well as the wider data science community. They will often work in multi-disciplinary teams with domain experts, Data Engineers, Business Analysts and Technology Professionals for example.

The EDISON Data Science Competence Framework (CF-DS) provides a methodological basis for designing the data science curriculum in universities and professional training institutions based on their Data Science Body of Knowledge (EDISON 2017a). The CF-DS provides a platform for consistency in definitions supporting the data science profession, across five groups of competences.

Three core competence groups correspond to Data Science related professional profiles as follows:

- **Data Analytics** – applies appropriate analytical/statistical techniques on available data to discover new relations and deliver insights to support decision-making
- **Data Science Engineering** – uses engineering principles and modern computer technologies to research, design and implement new data analytics applications
- **Domain-based Business Analytics** – adopts general data science methods to specific domain data types and presentations, data and process models, organisational roles and relations

And two additional common competence groups demanded by organisations include:

- **Data Management** – develops and implements strategy for data collection, storage, preservation and availability for further processing
- **Research Methods and Project Management** – creates new understandings and capabilities to discover approaches to achieve research or organisational goals

Following the ESCO (European Skills, Competences and Occupations) taxonomy, EDISON proposed extended Data Science professional profiles – three of these roles (data scientist, data analyst, business analyst) have been selected here to showcase a mapping with each of the five Data Science competence groups above. Figure 1 demonstrates graphically the extent of competence required for each job role, using the suggested ranking from EDISON (where 5 is highest relevancy).

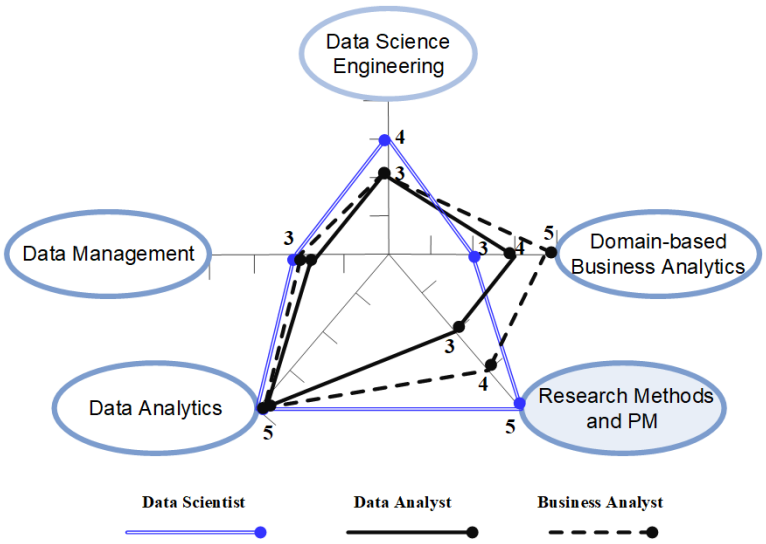

Figure 1: Mapping representative professional profiles with Data Science competence groups

No surprise to see that Data Analytics competence is essential to all three selected professionals (as shown at the bottom left of Figure 1). It is also interesting to note that Research Methods and Project Management is most relevant for Data Scientist (bottom right of Figure 1). Knowledge of research methods and techniques is especially important for data scientists and distinguishes them from other data-related professionals.

The Institute for Apprenticeships and Technical Education works with employers to shape skills training in England for the future of the workforce. An overview for Data Scientist has been described simply as “working in a team to find ways to improve an organisation's processes”. The IfATE standard has specified the areas of Knowledge data scientists must understand, the Skills they are able to deliver and the Behaviours they should demonstrate (IfATE 2018). For example, one required skill in relation to Research Methodology is defined as:

> *"S1. Identify and clarify problems an organisation faces, and reformulate them into Data Science problems. Devise solutions and make decisions in context by seeking feedback from stakeholders. Apply scientific methods through experiment design, measurement, hypothesis testing and delivery of results. Collaborate with colleagues to gather requirements."*

Not only are apprentices required to comprehend and reformulate the problems into Data Science tasks, but also apply scientific method throughout to design an experiment, test hypotheses and present results. At the highest level, data scientists are expected to put the problem in a wider context at their workplace, producing innovative methods and results that go beyond the original scope.

In terms of behaviour, apprentices must display scientific integrity and transparency, explaining the limitations of approach and results as well as the extent of their applicability to decisions in the organisation. For example: *"B5. An impartial, scientific, hypothesis-driven approach to work, rigorous data analysis methods, and integrity in presenting data and conclusions in a truthful and appropriate manner."*

The next section will illustrate the mapping of KSBs defined by the IfATE Data Scientist (integrated degree) standard with the body of knowledge and competency groups from EDISON, aiming to provide an organised and structured way to evaluate the knowledge and skills required for a Data Science apprenticeship.

3. Evaluation of KSBs with EDISON Framework

The University of Winchester has worked in collaboration with IfATE as well as a number of international, national and regional employers in the design and development of its Data Scientist (integrated degree) programme, which leads to a Bachelor's in Data Science. The BSc (Hons) Data Science degree apprenticeship supports students in gaining the knowledge and skills that are in demand by employers today and into the future, where working in multi-disciplinary teams alongside domain experts is often the norm.

The University has worked closely with its business partners to ensure that work-based training, research-informed teaching and rigorous academic study have been blended effectively to meet the needs of both the apprentice and employers (Lu 2020). This programme is built on activities which develop the different aspects of data science and leads to an Honours degree qualification that is contextualised for workplace competency as a data science professional capable of making a real contribution to their employer.

Degree apprenticeships offer employers the opportunity to develop and up-skill existing staff and to attract and train new talent. Some apprentices taking the Data Science programme are current employees already working in various roles, e.g. Business Insight Analyst, Claims Data Analyst, Commercial Pricing Analyst, Data & Analytics Manager, Management Consultant, Operational Researcher, People Insights Consultant, Product Manager and Underwriter. As a qualified degree apprentice, they will become an agile and skilled data scientist capable of rising to the challenges of the workplace. They will be prepared for a wide variety of professional roles in the field including not only data scientist but also business intelligence analyst, IT consultant, machine learning practitioner or IT project manager.

IfATE has designed the End-point Assessment (EPA) plan to enable the apprenticeship to be completed in accordance with its Data Scientist (integrated degree) standard (2018). Figure 2 sets out the context of EPA which includes three assessment methods: (1) Knowledge Test, followed by (2) Report (based on a work-based project) and (3) Professional Discussion (informed by a portfolio). The Data Science (Apprenticeship) programme design at Winchester has implemented EPA through two modules within the degree:

- **Data Science Project**: meets the requirements of "a suitable work-based project" which is "based on a Data Science solution that forms part of the apprentice's role" – the assessment of this module corresponds to EPA (2), i.e. *Report* – this should cover a significant work-based project detailing its aim and objectives, the apprentice's responsibilities and the action taken by them (in planning and executing the project)
- **Data Scientist Portfolio**: "enables the apprentice to demonstrate how they have applied their Knowledge, Skills and Behaviours in a real-work environment" – the assessment of this module corresponds to both EPA (1) and (3), i.e. *Knowledge Test* and *Professional Discussion* – the latter informed by a portfolio which presents evidence from other real-work projects

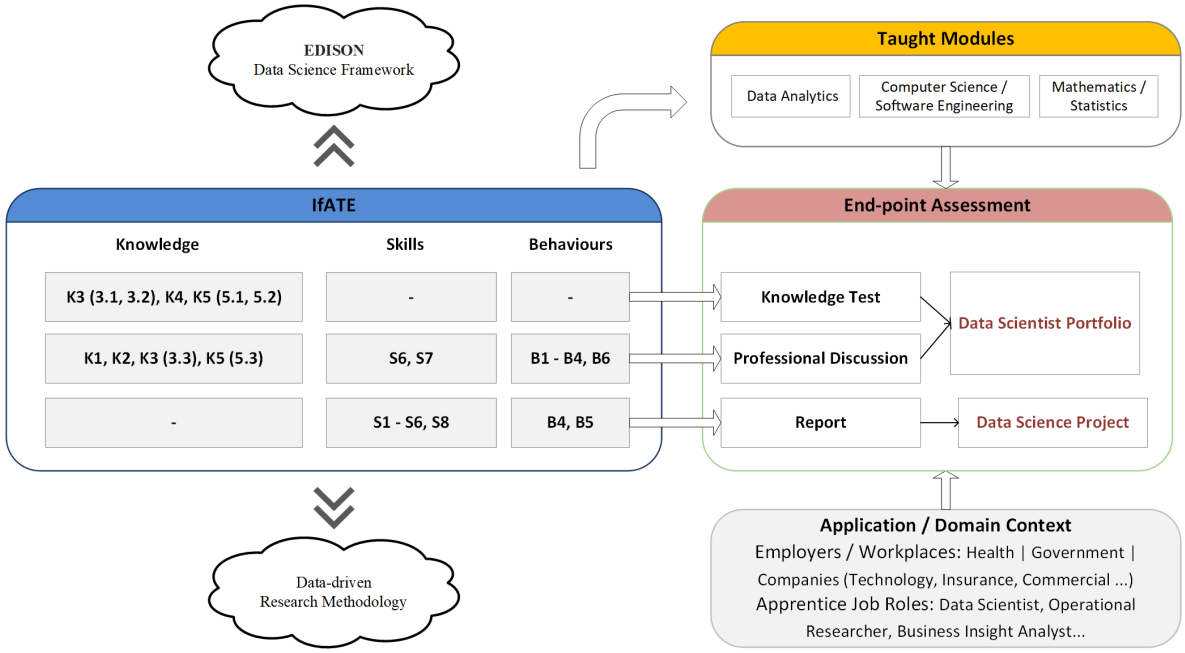

Figure 2: Context setting for End-point Assessment in relation to KSBs, academic modules and workplace

IfATE standard defines 16 areas of Knowledge, 8 Skills and 6 Behaviours which the University should follow through their teaching for Data Scientist (integrated degree) and these will also be tested through End-point Assessment. It is worthing pointing out that, as the name indicates, the Knowledge Test only assesses relevant knowledge, i.e. not skills or behaviours – on the other hand, the Data Science Project Report is not assessed for knowledge as such.

EDISON (2017b) proposed an extensive Data Science Body of Knowledge (DS-BoK) structure with identified Knowledge Areas in the same five competence groups mentioned in section 2, in summary:

- Data Analytics: including machine learning, data mining, statistical methods and visualisation
- Data Science Engineering: including software engineering, database and big data technologies
- Data Management: including data governance, data curation and preservation
- Research Methods and Project Management: including data-driven research and project quality
- Domain-based Business Analytics: including business analysis planning, processes and operations

DS-BoK (EDISON 2017b) suggested detailed data science knowledge areas, with corresponding knowledge units as well as certain alignment with other existing bodies of knowledge (e.g. DMBoK, BABoK, PMI-BoK, SWEBoK, ACM BoK etc.). Thus EDISON DS-BoK has been selected as a baseline to map KSBs from IfATE.

Figure 3 shows the mapping in a graphical way by adapting the EDISON framework illustrated in Lu (2022a) and using K2 as an example – the Business Analytics dimension is not considered in the mapping example here as degree apprentices are working in their own organisations across various domains.

The knowledge areas in EDISON DS-BoK which have not been considered specifically in IfATE KSBs are labelled in Figure 3 in italic and grey colour, e.g.

- Data Management group: Data infrastructure
- Data Science Engineering group: Agile development, Big data technologies
- Data Analytics group: Qualitative analytics, Prescriptive analytics
- Research Methods and PM group: Data collection, Use case analysis, Data stewardship

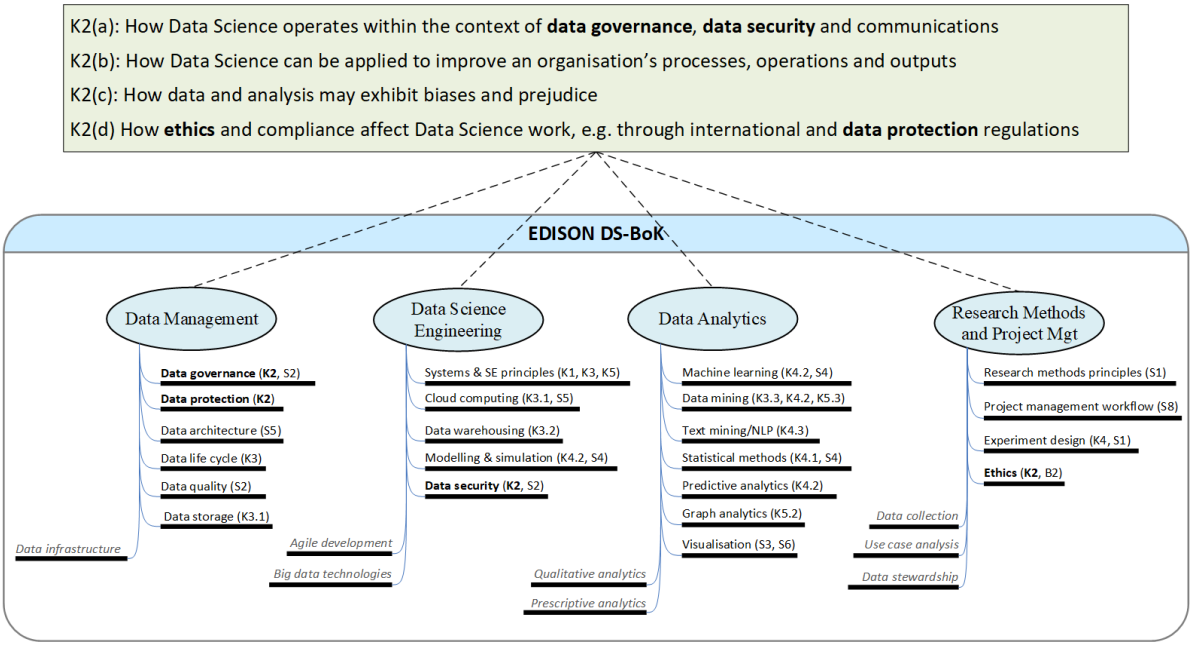

Figure 3: KSB mapping example in relation to EDISON framework

There are 35 instances of KSB in Figure 3, where 21 correspond to IfATE areas of knowledge, 13 to skills and just 1 to behaviour. Alongside the K2 highlighted, K4.2 also features in 4 of these instances, covering Modelling & simulation; Machine learning; Data mining; and Predictive analytics. At the same time, 13 of the 16 areas of knowledge are represented in the figure, 7 of the 8 skills and 1 of the 6 behaviours. The "missing" knowledge and skills so to speak typically belong to the more organisational business analytics domain. Behaviours can be considered among the softer skills and are not dealt with to the same extent in the more technological EDISON Body of Knowledge.

The next section will demonstrate the applicability of a data-driven research methodology in the context of the Data Science degree apprenticeship project as well as corresponding skills and behaviours from the IfATE standard.

4. Application of a Data-Driven Research Methodology

It is often necessary to develop a holistic approach to guide the process for executing a data science project (Martinez, Viles and Olaizola 2021). This section first reviews representative solution methodologies before presenting a data-driven research methodology linking the processes used in completing the EPA with relevant academic and technical skills. The section concludes by highlighting some sample real-work projects across different business and organisational domains.

4.1 Macro-Solution and Data Science Methodologies

Data science can be described as an interdisciplinary field of methods, processes, algorithms and systems to extract knowledge or insights from data, with data analytics as one of the key underpinning components. The Institute for Operations Research and the Management Sciences has produced their Analytics Body of Knowledge (Cochran 2019), which represents perspectives on a wide variety of analytics-related topics. INFORMS introduces the notion of micro- and macro- solution methodologies for analytics projects – macro-methodology provides the more general project path and structure, drawing on scientific research, OR/MS (Operations Research / Management Science) and CRISP-DM (Cross-Industry Standard Process for Data Mining).

For many years the most widely used approach for data analytics has been CRISP-DM (Provost and Fawcett 2013, Saltz 2021). This methodology breaks up the overall task of finding patterns from data into a set of well-defined sub-tasks: business understanding, data understanding, data preparation, modelling, evaluation and deployment. IBM Analytics has proposed a 10-stage Foundational Methodology for Data Science (IBM 2015) which extends CRISP-DM. It spans various technologies and approaches to form an iterative process for using data to uncover insights. The methodology has been redrawn here in the middle of Figure 4 to provide a central spine facilitating comparison with the scientific research and OR/MS project methods.

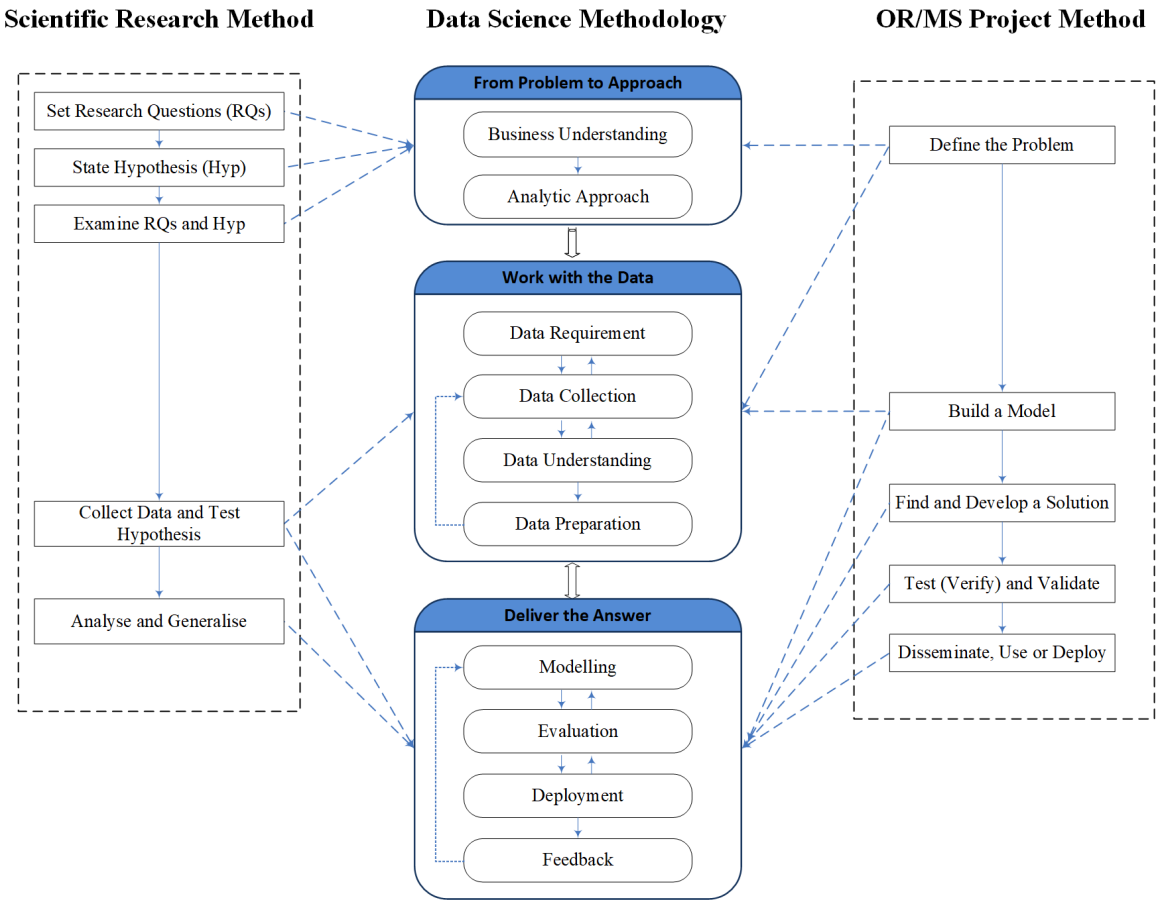

Figure 4: Comparing scientific and OR/MS project methods with a data science methodology

Saltz and Krasteva (2022) conducted a systematic review to explore the current approaches for executing data science projects and indicated that workflow is one of the most popular deployed with pre-defined phases, tasks and activities. There has been consistent research in data science process and methodologies, including the most recent work from Li et al. (2023) to embed data innovations in organisations by identifying data roles and associated skills with a workflow approach.

4.2 A Data-Driven Research Methodology Associated with KSBs

It is important to note that, while each data science project is unique, they can usually benefit from following a well-established methodology which recommends the steps to follow in a clear process (Cochran 2019). This will not only underpin credibility for the project path but also provide a structure for handling data and actionable insights. A composite data-driven process model was proposed previously and extended to provide both a procedural approach and structure which organisations can apply (Lu 2022b). This methodology has been evaluated in the business environment as a systematic process with well-structured stages. Case studies have included applying visual analytics on primary customer data to assist the strategic management of sales (Lu, Cairns and Smith 2021); as well as applying machine learning on a secondary health insurance dataset to gain knowledge about consumers and generate insights to remain price competitive (Lu 2022b).

The composite data-driven process model has been adapted in this paper as a consolidated data-driven research methodology which potentially can be used to guide a data science project for apprentices. As shown in Figure 5, the left-hand side indicates three distinct stages namely (i) Explore Data, (ii) Discover Insights and (iii) Make Decisions, each with its own structure. The extension in this paper is achieved by linking with an additional dimension, as shown on the right-hand side of the figure – the corresponding Skills and Behaviours which a work-based data science project report should demonstrate, according to the IfATE standard.

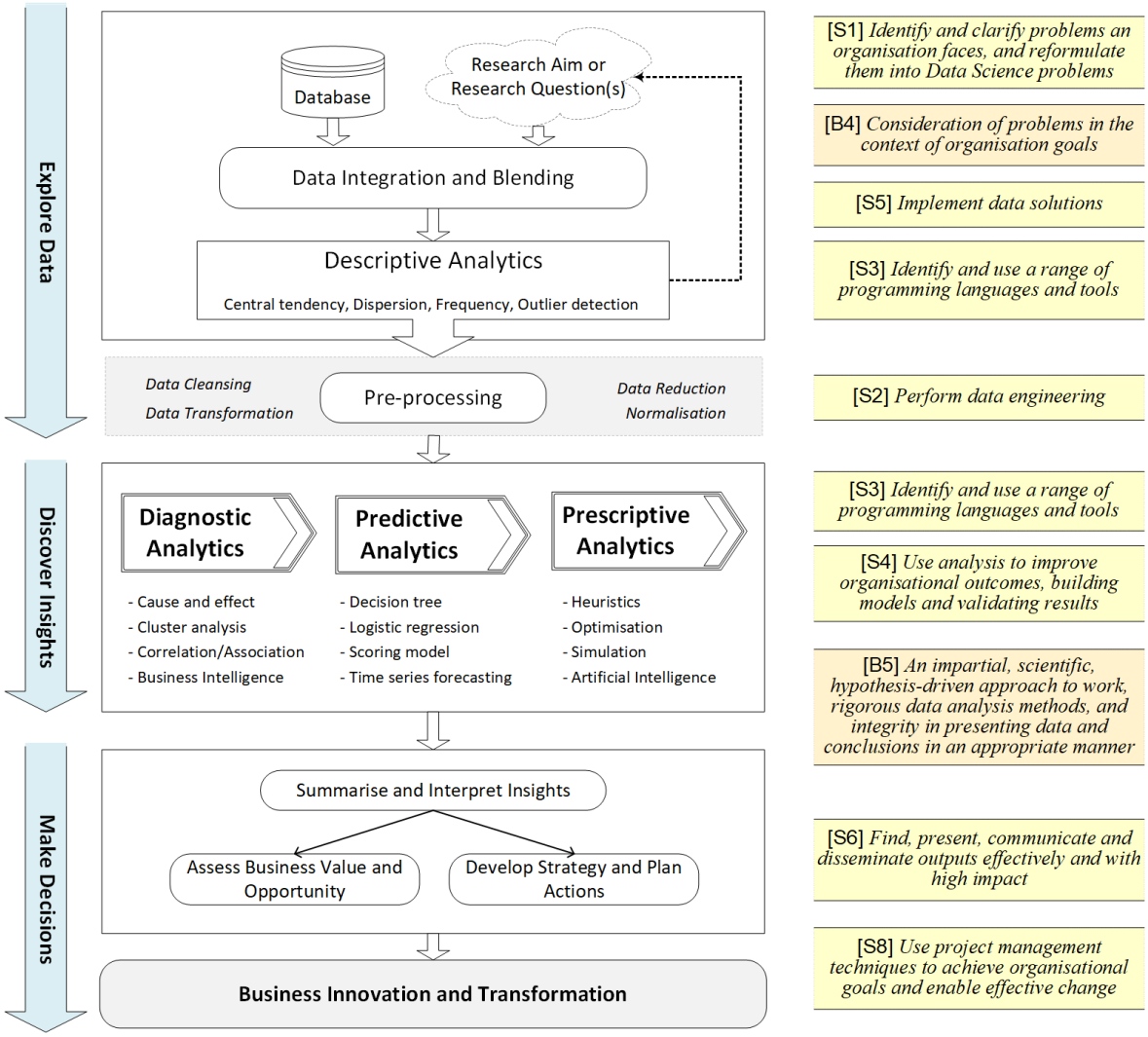

Figure 5: A data-driven research methodology with corresponding skills and behaviours defined by IfATE

The Data Science Project represents an ideal opportunity for the apprentices to showcase their Skills and Behaviours (SBs) as required by the Data Scientist degree apprenticeship standard. The project must be based on a Data Science solution that forms an integral part of the apprentice's role. The apprentice should show evidence of the planning and execution of the work-based project, including an evaluation of the processes followed, measurement of the SBs, a summary of learning derived from the execution of the work-based project as well as recommendations for future activities.

4.3 Selected Samples of Real-Work Projects

This paper will not evaluate in detail the application of data-driven research methodology in the real-world environment. However, some work-based data science projects from apprentices have been referenced in Table 1 by highlighting the domain-based context, datasets used, research aim (or questions) proposed and analytical methods applied.

These examples cover several domain-based business analytics areas including financial analytics, marketing analytics and human resource analytics – by using multiple data sources in the real-world across banking and insurance industries, sports as well as healthcare. Apart from Excel spreadsheets and SQL for data preparation, popular analytical tools include leading programming languages (Python and R) and visual analytics software (Tableau and PowerBI).

Table 1: Data, technologies and context – Examples of data science applications in the workplace

Context / Application and Datasets	Research Aim	Analytical Tools and Data Science Methods
Commercial Banking **Virtual assistant dataset**	To understand the causes of abandonment in customer conversations and identify how these may be grouped to improve helpfulness in the customer experience	Tableau (visual analytics) Python (sentiment analysis, clustering)
Health **Mental health dataset**	To investigate the impact of ethnicity, gender and age on mental health waiting times	SQL (data preparation) Python (regression model, time series analysis)
Insurance **Claims data**	To implement and test a method of Bayesian parameter estimation in commercial insurance modelling	Python Excel (numerical integration)
Sports **Football match data**	To find a correlation between the use of substitutes (subs) and total expected goals (xG) as an outcome variable	Python (multi-linear regression and decision tree)
HR **Employee data**	To allow stakeholders to view the diversity make-up of the business and to track progress against diversity goals	SQL server (querying data) PowerBI (dashboard)
Marketing **Leads and sales data**	Improving the comprehension of how diverse media channels influence the conversion of a sale, utilising the power of data science	R (time series analysis)

In terms of the Data Science Project Report assessment, in particular for the Methodology category, apprentices need to demonstrate:

- Are the methods which have been employed appropriate for the aim and objectives of the project?
- Have appropriate research methods been used?
- Is there evidence throughout the report that the methodology has been followed?
- Have the ethical implications been considered when discussing the methodology?
- Have the ethics of researching within the apprentice's own organisation been considered?

5. Conclusions

One of the key factors to determine the success of data science projects is the process followed – through established and mature methodologies (Saltz and Krasteva 2022). In this paper, the knowledge and skills required for a Data Science degree apprenticeship have been evaluated both from a UK standard and European perspective – data-driven processes are then linked to apprentice work-based projects and shown to be applicable through a research methodology.

The contribution of this paper lies in relating an occupational standard for Data Scientist with the still emerging professional practice through competence and skills mapping. Significant alignment is demonstrated between key aspects of the IfATE standard and EDISON framework – the former includes the more organisational and behavioural elements while the latter tends to emphasise the more technological. A consolidated data-driven research methodology then links the data exploration, knowledge discovery and decision-making stages of a work-based data science project with corresponding skills and behaviours specified by IfATE. This serves to underpin the domain-based knowledge and expertise required not only to achieve business and organisational goals, but also for successful completion of End-point Assessment.

The employer's role throughout the EPA exercise is integral to success too. They will not only continue their support and mentorship with the apprentices, but also participate in regular discussions with the University. The employer is required to offer necessary support with the Data Science Project, allowing the apprentice to carry out projects/tasks of significance to demonstrate their KSBs. As one of the key objectives for data science

projects is to offer business value, the data/information is likely to be confidential. As such the employer will need to approve of the project before submission, as well as the evidence within the portfolio, to ensure the level of information offered is appropriate.

The scope of data science "work" goes beyond analytics (Lahiri and Saltz 2023), not just being limited to the technical aspects such as modelling data and applying statistical or machine learning methods (McGregor and Inibhunu 2022). Data scientists can also have an impact at strategic and operational levels by building and maintaining collaborative relationships with key stakeholders, subject experts and colleagues as well as the wider data science community. The data-driven research methodology presented in this paper does not involve team collaborations at this stage, given its primary purpose of guiding an individual data science project which can then be assessed.

As part of the future work, further "soft skills" for data scientists could be embedded within the research methodology in order to accommodate other behaviours required by IfATE, notably: "*B2. Empathy and positive engagement to enable working and collaborating in multi-disciplinary teams, championing and highlighting ethics and diversity in data work"* as well as *"B6. A commitment to keeping up to date with current thinking and maintaining personal development. Including collaborating with the data science community"*. Employers offer additional support through a testimonial to acknowledge demonstration of the behaviours, as it can be difficult to evidence these through documentation for the EPA – it is also an opportunity to focus on the apprentice's competence from a management perspective.

References

Cao, L.B. (2017) "Data Science: Challenges and Directions", *Communications of the ACM*, Vol. 60, No. 8, pp.59-68.

Cochran, J.J. (2019) *INFORMS Analytics Body of Knowledge*, John Wiley & Sons, Hoboken, NJ.

EDISON Project (2017a) "Building the Data Science Profession", [online], https://edison-project.eu/edison/edison-data-science-framework-edsf/

EDISON Project (2017b) "Data Science Body of Knowledge (DS-BoK)", [online], https://edison-project.eu/data-science-body-knowledge-ds-bok/

IBM Analytics White Paper (2015) "Foundational Methodology for Data Science", [online], https://tdwi.org/~/media/64511A895D86457E964174EDC5C4C7B1.PDF

IfATE (2018) "Institute for Apprenticeships and Technical Education: Data Scientist (Integrated Degree)", [online], https://www.instituteforapprenticeships.org/apprenticeship-standards/data-scientist-integrated-degree-v1-0

Lahiri, S. and Saltz, J.S. (2023) "The Need for an Enhanced Process Methodology for Ethical Data Science Projects", IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS), West Lafayette, IN, USA, pp. 01-05.

Li, K., Griffin, M.A., Barker, T. et al. (2023) "Embedding Data Science Innovations in Organizations: A New Workflow Approach", *Data-Centric Engineering*, Vol. 4, No. 26.

Lu, J. (2020) "Data Analytics Research-informed Teaching in a Digital Technologies Curriculum", *INFORMS Transactions on Education*, Vol. 20, No. 2, pp.57-72.

Lu, J., Cairns, L. and Smith, L. (2021) "Data Science in the Business Environment: Customer Analytics Case Studies in SMEs", *Journal of Modelling in Management*, Vol. 16, No. 2, pp.689-713.

Lu, J. (2022a) "Data Science in the Business Environment: Insight Management for an Executive MBA", *The International Journal of Management Education*, Vol. 20, No. 1, Elsevier.

Lu, J. (2022b) "Data Science in the Business Environment: Architecture, Process and Tools", In: D. Garg et al. (Eds.) *Advanced Computing. Communications in Computer and Information Science*, Vol. 1528, Springer Cham, pp.279-293.

Martinez, I., Viles, E. and Olaizola, I.G. (2021) "Data Science Methodologies: Current Challenges and Future Approaches", *Big Data Research*, Vol. 24, No. 100183.

McGregor, C. and Inibhunu, C. (2022) "A Framework for the Design, Development, Testing and Deployment of Reliable Big Data Platforms", IEEE International Conference on Big Data (Big Data), Osaka, Japan, pp. 2660-2666.

Provost, F. and Fawcett, T. (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking*, O'Reilly Media, California.

Saltz, J.S. (2021) "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps", IEEE International Conference on Big Data (Big Data), Orlando, USA, pp. 2337-2344.

Saltz, J. S. and Krasteva, I. (2022) "Current Approaches for Executing Big Data Science Projects: A Systematic Literature Review", *PeerJ Computer Science*, Vol. 8, No. 862.