Training Data Scientists through Project-based Learning

Fernando Martínez-Plumed and José Hernández-Orallo

***Abstract*—The concepts of innovation, creativity, problem solving, effective communication, autonomy and critical thinking are at the core of becoming a good data scientist. Adapting to new technological resources and tools is also an important skill, which also builds on the curious and inquisitive nature associated with data science, and is fuelled by rapidly changing data science ecosystems in industry. In this regard, Project-based learning (PBL) has clear benefits for engaging students in data science courses. However, the exploratory character of data science projects, which do not start with a clear specification of what to do, but some data to analyse, pose some challenges to the application of PBL. Our aim is to improve students' data science learning experiences and outcomes through the use of PBL. In this paper, we share our experiences with PBL and present an assessment rubric that focuses on value, innovation and narrative, which can be used as a scaffolding structure for data science courses. Our analysis of a PBL data science course at MSc level, together with data from student surveys, shows how the methodology and rubric align well with the exploratory nature of data science and the proactive, curious, and inquisitive skills required of data scientists.**

***Index Terms*—Data Science, Project Based-Learning, Assessment Tools.**

I. Introduction

THE search for new teaching methodologies in higher education and lifelong learning is a widely debated topic at universities and other training environments. The ease of access to information (as well as the many and varied ways of obtaining and contrasting it) have influenced the profiles of younger and more mature students, at physical or online courses. Furthermore, in contrast to encyclopaedic knowledge, a person today can hardly ever master all the knowledge in a very specific field. The accelerating generation of new knowledge urges us to equip the educational and training systems with new learning techniques that make all stakeholders, including instructors, students and employers, more suitable for this process of continuous change. This motivates an open debate around the search for new methodologies that make students learn more effectively, with the aim of training professionals adapted to this new society. New models look for learning process that are more student-centred than teacher-centred. For this to happen, teaching methodologies have to change. The so-called active methodologies [1] play a preponderant role in achieving this objective.

One major active methodology is Project Based Learning (PBL), a cooperative learning strategy that understands learning as a communication process and focuses on the learner as both an individual and a member of a group. In PBL, solving a problem, the project, drives the whole process [2] and the acquisition of the skills. Students are responsible for their own progress and teachers take the role of providing materials, feedback and support when requested, on top of being advisors to facilitate the students' work. Students are also able to immediately see their project as a constant and efficient testing ground for new ideas. Additionally, students are much more likely to understand and apply concepts if they can use their knowledge to effect change in the real world.

In the area of computer science, PBL has been proved to be one of the most engaging elements for students [3]. Many software engineering courses were pioneers using PBL, but their use has been extended to other subjects in computing and engineering more broadly [4]–[8]. In the particular character of data science projects we deal with in this paper, there are specific elements that require a well-thought combination of data science and PBL methodologies for the design of an effective PBL course. In an explorative data science project, there is no initial specification, unlike other PBL-based computer science or software engineering courses. Many data science projects even lack a clear goal, unlike directed data mining projects, which starts with a business goal that has to transformed into a data mining goal. In particular, in data science the *data* take centre stage: we know or suspect there is value in these data, how do we discover it? What are the possible operations we can apply to the data to unlock and utilise their value? While moving away from the process, the methodology we should follow when addressing a data science task becomes less prescriptive and more inquisitive: things you *can extract from* the data rather than things you *should do to data*. The key difference we perceive between the old, but related term, *data mining* twenty years ago and data science today is that the former is goal-driven and concentrates on the process, while the latter is data-driven and exploratory [9].

In directed data mining, a whole project can follow a sequence of stages starting from the 'business goal', translated to a 'data mining goal', which leads the rest of the process [10]. Accordingly, de facto methodologies such as CRISP-DM [11], the CRoss-Industry Standard Process for Data Mining, were conceived to catalogue and guide the most common steps in data mining projects. However, in data science, context becomes more relevant during the whole process. Accordingly, new processes challenging CRISP-DM have been introduced, by including context adaptation and model reuse [12], or proposing a more flexible view of data science projects as traversing trajectories [9]. Under this more accurate view of data science, the paths that a project can take become more varied and the order of activities depends on the domain as

F. Martínez and J. Hernández-Orallo are with the VRAIN institute, Universitat Politècnica de València, Spain (e-mail: {fmartinez,jorallo}@dsic.upv.es)

well as on the decisions and discoveries of the data scientist. The illustrative Figure 1 shows a space of exploratory, goal-driven and data-management activities which may —or may not— be performed in a particular data science project by following different trajectories in an order that is not predetermined. Many data science projects do not start from a clear specification, nor can elicit the specification from a client or expert, but the very value of the data is a journey, especially looking for insights and novelty. This journey ends up in a story, a data narrative, which has to be emphasised through an appropriate presentation and exposition. This exploratory journey of data science and the relevance of finding novel insights that bring value to a particular domain suggest that some small tweaks to the PBL methodology will not suffice. Instead, we need an important overhaul of the methodology and the associated procedures, most especially those rubrics that assess the skills required by data scientists.

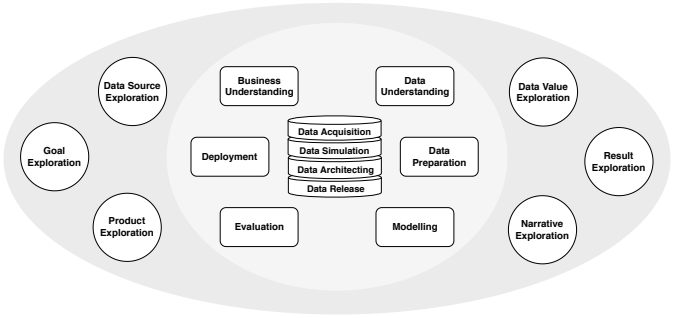

Fig. 1. The Data Science Trajectories map (from [9]), containing the outer circle of exploratory activities, inner circle of Data Mining (or goal-driven) activities, and at the core the data management activities.

Despite the rapid growth in the need for data science skills for many years, leading to a rise in the number of data science programs [13], little has been written about how to best educate data scientists [14]–[17]. Instead, we can find different generic pedagogical models such as flipped teaching [18], the use of methods and tools to support learning [19], or the consideration of collaborative efforts among instructors with different academic disciplines [20] (trying to increase participation and engagement [21], [22] and creativity [17]). Less commonly, we also find works analysing the importance of embedding a substantial practicum in the curriculum of data science courses (see, e.g., [23], [24]). The main novelty of this article is, therefore, to present a customised PBL-based methodology for data science courses whose main aim is to motivate students to improve skills such as innovation and creativity, problem solving, effective communication, autonomy and critical thinking, also promoting collaborative work. To this end, we make the following contributions:

- We describe a reframed teaching approach towards PBL in a data science course with differentiated real-world projects, also addressing several teaching aspects for the enhancement of the learning process and its evaluation.
- We bring innovative didactic resources that instructors can use to organise their courses, including a special rubric to assess the balance between risk-innovation vs. outcome for the project, which is used by the students as scaffolding.
- We illustrate the operation of the PBL methodology and use of the new rubrics presented here through the results of a case study in a MSc data science course.
- We evaluate the success of the teaching methodology through a questionnaire with rating scales (Likert-type) and open answers, including questions on motivation, objectives and achievement of competences.

The surveys show that more than 85% of the students indicate a high motivation and that both the objectives and the competences met their expectations.

The rest of the paper is structured as follows. Section II explains the new skills that data science profiles require. Sections III and IV present a running case study and the tailoring of the PBL for a data science course. Section V presents the project-based assessment tools and rubrics. Section VI analyses some illustrative examples. Finally, section VII discusses the evaluation of the methodology and the students' results, closing the paper in Section VIII.

II. Project-based learning and data science

The role of the traditional teacher who taught a subject through master classes and then, at the end of the course, gave an exam on the subject has been questioned on many occasions, with new models being sought since the late 1960s [25]. New teaching models reduced the role of lectures and developed students' creative abilities by posing questions and open-ended problems. The development of PBL began to be applied at the university level in the field of medicine [26]. Later, it was applied in engineering and specifically in computer science [27], [28]. Actually, the teaching of computer science disciplines proves to be a highly conducive framework for the implementation of project-oriented activities and subjects [27]. In this scheme, teachers propose one or several projects, usually inspired by real problems, which students must solve in groups. Students have to decide how to tackle the projects and which activities to carry out. This type of learning can be of great value in fostering the development of generic skills such as [29]–[31]:

- **Teamwork**: Working in teams develops coordination, communication, responsibility and planning skills, etc.
- **Resourcefulness**: The students' motivation is encouraged by the search for and understanding of new information, using all the resources available to them.
- **Proactivity**: The student is not a recipient of knowledge, but an active agent in their learning and problem solving.
- **Innovation and creativity**: Conforming to the norm is not a guarantee of success. Rather, innovative ideas are encouraged and allow them to stand out over the rest.
- **Abstract thinking**: PBL facilitates interdisciplinarity and higher-order thinking.
- **Formative and non-punitive assessment**: The aim of the assessment is for the student to learn from mistakes, which provides a richer learning experience.
- **Critical thinking**: PBL confronts students with real-world situations and they have to compare pros and cons for each single decision along the way.

Data Science is a fairly young field of science and technology —yet old roots in statistics and computer science— which is particularly well-suited for the adoption of PBL methodologies in teaching and evaluation. In brief, data science is an interdisciplinary field that involves scientific methods, processes and systems for better understanding of data in its various forms, whether structured or unstructured, extracting knowledge and deploying it in a variety of domain. It is actually the modern evolution of some other disciplines dealing with and analysing data, such as statistics, data mining, database technology, machine learning, and predictive analytics for the purpose of understanding and analysing real phenomena [32].

Over the past few decades, data science has increasingly become popular, becoming the part and parcel of every business model. While the classical area of data mining for deriving value from data has grown exponentially in size and complexity, it has also become much more exploratory under the umbrella of data science. In the latter, data-driven and knowledge-driven stages interact, in contrast to the traditional data mining process, starting from precise business goals that translate into a clear data mining task, which ultimately converts "data to knowledge". In other words, not only has the nature of the data changed but also the processes for extracting value from it.

It is then expected that this new profile of applications and projects would require new skills as well as the consolidation of data science as a new profession. Data scientists are expected to cover a wide range of soft skills, such as being proactive, curious and inquisitive, as well as being capable of communicating results, leading a team, being creative, etc. [33]–[36]. Most of the new exploratory steps imply soft skills. Also, the understanding of new domains must play an interactive and exploratory role in most data science projects [9]. It is no surprise that the more flexible, less systematic, character of the new exploration and data management activities (see Figure 1) highlights the challenges that characterise data science and paves the way for following innovation-based PBL methodologies when training data scientists.

III. Running case study

In the past academic years, the authors have adapted and tailored the PBL methodology for their data science course (CDA). This was motivated by the experience gained from other similar courses where the authors had seen that the application of classical teaching methodologies (e.g., lectures, laboratory practicals and independent work of the students) may not be the most appropriate for the competences of a data science course. With the aim of improving the learning and performance of the students, by increasing motivation and participation, the authors prepared this course paying special attention to the following questions: Will the students learning something useful? To what extent will they learn original solutions? And, most importantly, what would happen once they go out into the labour market and have to deal with real problems? Driven by these questions, we designed various alternative assessment scenarios to evaluate our students' performance based on the development of open-goal projects that would be conducted collaboratively (maximising utility, passion, curiosity and competition), inspired by the real-world project-based nature of data science.

CDA is an optional subject (taught in English) in the Computer Science Master degree at the School of Computer Science and Engineering (ETSInf) of the Technical University of Valencia (UPV), Spain. It is a four-month course and has been assigned 6 ECTS[1] credits distributed in 1.5 credits of classroom theory, 3 credits of seminar and 1.5 credits of laboratory practice.

As an optional course, since its creation in the academic year 2016/17, CDA has had around 30 students each year. Students come from different Computer Science programs from the UPV and other European universities (61% of the student in year 2021/22 were from outside the UPV). This course focuses on preparing students for the role of a data scientist in an organisation, so that they can identify data-related problems and opportunities, and deploy and communicate data-driven products using effective tools. The main goals can be summarised as:

O1: Recognise the value of data and the business opportunities for the development of data-based products, in the context of Big Data.

O2: Determine the technologies that are needed to handle data efficiently in different environments, different sizes and formats, to ease data understanding and analysis.

O3: Estimate the complexity and resources that are needed for a data analysis project and establish the measures of cost and success.

O4: Convey the results, implications and value of the analysis, building effective visual representation.

In addition, the competences and skills through which to achieve these objectives are as follows:

C1: Possess and understand knowledge that provides a basis or opportunity for originality in the development and/or application of ideas, often in a research context.

C2: Apply acquired knowledge and problem solving in new or unfamiliar environments within broader, multidisciplinary contexts, being able to integrate this knowledge.

C3: Integrate knowledge and face the complexity of formulating judgements based on incomplete or limited information, including reflections on the social and ethical responsibilities linked to the application of their knowledge and judgements.

C4: Communicate their conclusions —and ultimately knowledge and rationale behind them— to specialist and non-specialist audiences in a clear and unambiguous way.

C5: Possess the learning skills to enable further study in a way that will be largely self-directed or autonomous.

C6: Understand and apply the ethical responsibility, legislation and professional ethics of the activity of the profession of Computer Science.

C7: Integrate technologies, applications, services and systems specific to Computer Science, with a generalist character,

[1]The European Credit Transfer and Accumulation System (ECTS) is a standard means for comparing academic credits for higher education across the European Union and other collaborating European countries.

and in broader and multidisciplinary contexts.

To meet these goals through competences, the course emphasises the value of data and the role of the "data scientist" in real cases, with different kinds of data being integrated and manipulated using data science tools. Furthermore, students acquire further theoretical knowledge for the purpose of independently addressing the different activities of a data science project (e.g., data value exploration, exploratory data analysis, data cleansing and transformation, modelling, assessment, product exploration, etc.) but they are not usually able to apply it towards a common goal. It is then important for students to face an approximation of what a real (engineering or data science) project is: the synthesis of a product (within a team) from its conception to its delivery. Throughout the course, the objectives will be completed through the work carried out by the students, requiring or developing different competences.

On the other hand, the course also promotes and develops generic skills or transversal competences (TC) [37]. These are skills not specifically related to a particular area of knowledge, but that can be used in a wide variety of situations and provide competitive advantages for students entering the job market. In particular, our data science course develops the following skills:

TC1: Effective communication: Communicating effectively means having developed the ability to transmit knowledge and express ideas and arguments clearly, rigorously and convincingly, both orally and in writing, using resources appropriately and adapting to the circumstances and type of audience.

TC2: Innovation and creativity: The development of this competence requires both thinking differently in order to provide different perspectives (creativity) and committing certain resources on one's own initiative in order to explore an opportunity, assuming the risk that this entails (entrepreneurship).

TC3: Analysis and problem solving: This competence refers to the need for students to be able to apply structured procedures to solve problems and to make decisions, thus promoting their ability to learn, understand and apply knowledge autonomously, as well as to understand the mechanisms of knowledge expansion and dissemination.

IV. Data Science Project Definition

Aiming at adopting a PBL methodology in CDA, we must carefully choose the projects to be developed and, specifically, what the main characteristics of these data-oriented, exploratory projects should be, also trying to unlock the prescriptive and inquisitive skills of the students. First, it was first necessary to establish what kind of projects to solve. To increase motivation and to improve students' innovation and curiosity skills, we decided that it should be real-world problems so that the students could come up with novel contributions in a given field. However, the main challenge (and risk) is that the project proposal is conceived as a freelance data scientist project and, thus, the students are responsible for the project definition and development. No clear specifications, guidelines, or recommendations are to be provided to the students, unlike other PBL-based computer science or software engineering courses. They are also not provided with any sort of predefined topic, stages, template or technology to be used. The reason behind this is that, if students suggest the project and how to solve it, their motivation is very high [38]. This way, the students themselves have to develop the idea of a new product from data, or the improvement of an existing procedure with data-acquired knowledge. Students would choose to develop data-based products on a topic they are genuinely curious about, also thinking on what an employer or the general public would want to see. In any case, the instructors must provide some guidance and apply certain techniques to keep the students motivated, such as helping the students with setting up their projects, selecting the topics that may be more attractive, providing open data repositories, etc., as well as encouraging healthy competition among them if the projects to be developed are related.

Data-oriented projects should also be end-to-end, requiring students to form the teams, explore the options for data value, perform the market research, specify the goals, identify the data sources, design solutions, perform exploratory data analysis, build, evaluate, deploy and maintain models, and deal with the communication and presentation issues if applicable. Students will manage their schedule as well, although some guidelines are provided along with the deadlines for the different phases to develop, such as team formation, market research, data gathering, implementation, final deployment, reporting, etc.). During the development of the project, the instructor plays the role of an "advisor" (a guidance counsellor) meant to provide assistance whenever necessary, mainly in the development of the different phases of the projects (e.g., help in the search for new ideas, provide useful sources of data, answer technical questions, etc.). From a pedagogical point of view, there are a number of "tasks" that the instructor must perform for the project to be successful:

- **Seminars**: Seminars will serve to introduce tools, methods, etc., that the student is unfamiliar with and that may be somewhat difficult to learn from the beginning in a self-taught way. Afterwards, it will be the students who will go into them in depth on their own.
- **Group monitoring**: The instructor must know the evolution shown by each group in the solution of the project and the degree of involvement of each student in it (not all students learn the same or in the same way). Effective monitoring can be done during the classes, or in a more personalised way through groupwise meetings, and it depends on gathering information and giving feedback on groups interactions, as well as anticipating and preparing for potential problems (e.g., inadequate progress, members not contributing, etc.).
- **Group feedback**: To support students in moving forward with their project, the instructors must constantly provide qualitative and/or quantitative feedback to reduce the gap between actual and expected learning outcomes. Feedback is provided following two mechanisms: 1) face-to-face discussions where specific issues of the projects are discussed; or 2) formal written feedback. We see this

TABLE I
ASSESSMENT RUBRIC USED IN PBL-BASED DATA SCIENCE COURSES. ITEMS ARE WEIGHTED EQUALLY.

Item	**Excellent** ($\simeq 10$)	**Good** ($\simeq 7.5$)	**Satisfactory** ($\simeq 5$)	**Needs improvement** ($\simeq 2.5$)
Data VALUE **Weight**: 20%	The students have identified the value of the data they have worked with, their applicability in their contemporary world, the people who may be benefited by this work, and possible apps or even a future entrepreneurship idea as an outcome. They have also identified the limitations and sustainability issues, as well as the overall impact of the use of the data and the proposed idea on society.	The students have identified the value of the data and its applicability, the beneficiaries and final limitations. They also identify the impact of the data product and the risks of its use.	The students have briefly identified the final value of the data and its applicability. The project is sometimes dominated by the details without seeing the big picture.	The students describe the data, but fail to transmit what the purpose of all this will be and why this will provide value and why it will be novel.
ALTERNATIVES and innovation **Weight**: 20%	The students have looked for alternative proposals (bibliography, websites, apps) for the same domain, the same data or application, and have compared (quantitatively) their results with them at least at the abstract level), and seen whether what they present is the same or innovative, is below the current state of the art, covers real needs, etc. Preliminary market studies are well received.	The students have looked for some alternative proposals for the same domain, data or application, also making some qualitative or abstract comparisons, pointing to the innovations of the project in general terms.	The students have looked for few alternative proposals in a general way, with little or no relation to the presented project. Very generic comparisons are made, and innovations are not presented in a clear and specific way.	The students have overlooked alternative proposals and no comparisons are made.
TECHNICAL tool integration **Weight**: 20%	The students have mastered different new technical tools (e.g., development IDEs and notebooks, Python/R data analytic and modelling libraries, visualisation software, data/web scrapping, API management, etc.) and their integration to appropriately meet their goals. The work and expertise seen during the course are reflected in the technical solutions, which take the best option from the state of the art and the literature. The solutions show initiative and originality.	The students have integrated several tools (e.g., Python/R data analytic, modelling and visualisation libraries, data scraping, etc.), and used them appropriately. The solutions reflect an important amount of effort and adequacy for their needs.	The students have used some tools (e.g., basic libraries and software) appropriately for their goals. The work and expertise seen during the course are reflected in the technical solutions.	The students use inappropriate tools. Not enough effort has been put in finding the right tools or learning new ones.
Project EFFORT **Weight**: 20%	The students have worked with different data repositories, and they have made a major effort of curation, integration and collection (e.g., through appropriate ETL data integration process). Also, they tried many different models and variations of their features (e.g., through tuning grids and hyperparameter search), chosen the right metrics and evaluation protocols (e.g, split, hold-out, k-fold, etc. validation recipes). Students show the lessons learnt and how they have solved the problems or found a way around. They show a clear evidence of teamwork.	The students have used, curated and integrated different data repositories (with some integration effort through appropriate ETL processes), tried different models, chosen the right metrics and evaluation protocols. They show some lessons learnt.	The students have used few data repositories (with little integration), and tried a few models. There are some issues in the evaluation of the models, the chosen metrics or the lessons learnt provided.	The students have used very few data repositories (with little or no integration), tried very few models and/or evaluated them inappropriately, with wrong model selection or overfitting, and drawing wrong conclusions.
EXPOSITION quality **Weight**: 20%	The students have been able to transmit the ideas very clearly, the motivation of the work and the insights. The quality of the slides and the graphics are impeccable, with the right element for illustrating each point of the story. They make gestures, use expression resources and really engage with the audience. They are telling a story. They answer the questions correctly and precisely.	The students have been able to transmit the main ideas and the results. The presentation is well organised and supported by graphics. They make some gestures to avoid being monotonous. They are telling a story. They answer the questions correctly.	The students transmit what their project is about. The presentation does not seem to have a clear organisation, and graphics are used by availability rather than opportunity. They make a monotonous presentation. They answer most of the questions correctly.	The students are not able to transmit the key ideas of their project. The presentation is messy and not well supported by graphics. They are boring, make many mistakes or do not know how to follow. They answer many questions wrong or vaguely.

in further detail in the next section.

V. PROJECT ASSESSMENT

The grading of the subject is based on three evaluation acts: 1) Short in-class quizzes (two assessments, 10% of the final mark each); 2) practical assessments (three assessments, 10% of the final mark each); and 3) a work (project) including oral presentation (50% of the final mark), the latter being the main component of PBL. To assess the performance of the students within these projects, the students create a project portfolio defining the activities, which is also used for assessing the students work. The project has to be carried out in groups of 2-4 people. Since the students have to work in a team, they have to employ transversal competences such as flexibility, organisation, problem solving, negotiation skills, leadership, etc.

All projects require a final presentation (carried out by all members of the group) describing the main development and results. It is worth mentioning that a very common mistake is to grade this sort of projects from an excessively utilitarian point of view. The project product is not intended to be directly usable by society. While the project must deliver a product, the evaluation should focus on *how resourceful the students are during the process*.The fact that there are "problems" the students detect and solve, should be encouraged and not discouraged, and so should be reflected in the grading.

Students will be able to ask for feedback during the development of the project. Also, students have a (first) pre-assessment (rehearsal) of their presentation, receiving point-by-point feedback. Two weeks after this first attempt, they can do the (second) final presentation/assessment. The pre-assessment grade can be considered final if students decide not to resubmit their work for the final assessment. There is no further resit after this final assessment. This project rehearsal-improvement-resit process takes place in two weeks. During the presentations, students from other groups should ask questions and express what the project conveys to them. The instructors write a detailed report following an assessment rubric (shown in Table I). This first evaluation comes with a provisional mark. Based on this feedback, which includes the different sections of the rubric, especially the presentation, the group works on improving the scope of the project and can make a final presentation in the last week.

The goal of the rubric is not only to make assessment more systematic. It is also a guide of what is going to be valued, and serves as an abstract scaffolding for the students about where they should put their focus and effort. For the development of the rubric, we have tried to identify those target standards or skills for data scientists [39], [40] that will frame and focus

their work, as well as to identify the essential criteria we would like to assess. The rubric should thus help assess all multi-disciplinary knowledge and competences that are required from the data science practitioners in data driven projects [15], [20]. In this regard, the rubric seeks to emphasise a meaningful set of skills and different aspects linked to the exploratory nature of data science, such as the discovery of the value that might be extracted from the data, the novelty and challenge of the approach, the exploration and preparation of the data, and its subsequent model building and evaluation, ending with a product or study. The exposition of results (written or oral) must be accompanied by a well-thought narrative to attract the audience to the final product or understand the key insights. At the technical level, the rubric will be used to assess the learning outcomes of students in terms of their ability to use (open-source) tools and methods to collect, process, store and analyse data; apply machine learning and data mining techniques to generate interesting business insights from the processed data; and effectively test hypotheses and interpret predictive models applied to the data.

Also, the rubric allows for the assessment of those transversal competences developed in the course through the different items it includes. In this regard, the effective communication (CT1), as a process of purposeful exchange and presentation of ideas, thoughts, knowledge and information, is evaluated through the item *exposition quality* in the Rubric (Table I). The item *alternatives and innovation* serves to measure the originality of the ideas and whether the project has some viability, which corresponds with CT2. Finally, problem-solving the analytical and problem solving skills (CT3) is evaluated through the item *technical tool integration*.

TABLE II
GROUP CO-EVALUATION RUBRIC USED IN OUR PBL COURSE.

Item	Description
Contribution	What's the percentage of the total contribution that can be attributed to your teammate X? (considering the result of the project, not hours of work, as some people are more efficient than others)
Disposition	On a scale between 0 and 100, how would you value the collaborative attitude of your teammate X? (disposition, helpfulness, seeking consensus rather than conflicts, etc.)

Furthermore,we decided that it was necessary to also adopt cross-assessment techniques to better assess the performance of each student, so that the grade is more in line with the contribution of each student within the group. In this regard, the project assessment requires every student to complete peer evaluation about the performance of their group members, as shown in Table II. This evaluation is carried out after the presentation of the project. The sum of "contribution" for all participants in a particular team should be 100. Given the above pair of marks for all team members in a group, teachers use a procedure (not disclosed to avoid optimising for it instead of being honest), to derive a coefficient between 0 and 1.2 for each student, according to the values and harmony of these cross-assessments. This coefficient should be close to 1 if the student has made a fair contribution to the share of the project, has shown a collaborative attitude towards their teammates, etc. Finally, this coefficient multiplies the score given by the instructor to obtain the final individual grade for the project.

VI. PROJECT EXAMPLES

For illustrative purposes, we show the feedback provided (via email) for a couple of projects with different levels of maturity. The first one proposed a machine learning approach to analyse and predict employment and its related economic factors based on data from the stock market (IBEX35). For the first assessment, the project was still at a preliminary stage and suffered from a lack of a clear motivation and justification, poor narrative, nonexistent comparison with related work and insufficient coherence and cohesion regarding the analysis and results. After the feedback (see Table III), in the resit, the group improved their final (average) score from 2.8/10 to 5/10, addressing some of the concerns raised.

TABLE III
FEEDBACK PROVIDED FOR AN IMMATURE PROJECT ON THE ANALYSIS OF THE RELATION OF EMPLOYMENT AND OTHER SOCIAL INDICATORS BASED ON DATA FROM THE STOCK MARKET.

Item	Feedback
Data VALUE **Score**: 3/10	**Well done**: Employment is an important problem and four big companies may have an effect on employment. **Can be improved**: Not clear what value is to be found here. Companies can increase benefits and reduce employees or the other way round. But the relationship is complex and using four companies is going to be inconclusive anyway. We would need a clear case of an important question that is to be answered, such as those the students introduce in the conclusions, and who is going to be benefited by the insights of this study and how.
ALTERNATIVES and innovation **Score**: 0/10	**Well done**: - **Can be improved**: There's enormous economic literature about the effect of employment and the trends of companies, but typically the studies use economic models and more data. I suggest the students to consider some data that is not usually analysed in economic studies, to make this is a little bit more innovative and not compete with expert economists. It must be something different to make a point.
TECHNICAL tool integration **Score**: 3/10	**Well done**: There's some processing reading the CSV and integrating and preparing data. **Can be improved**: The analysis is merely a collection of bar plots and line charts. We need to use some of the modelling tools to find clusters, trends, sequences, etc. This may require more data than just four companies, or more detailed information for each company.
Project EFFORT **Score**: 5/10	**Well done**: There's some effort fetching and integrating the data. **Can be improved**: There's insufficient data and insufficient analysis. More effort should be done in many other aspects of the project, including value, novelty, modelling and presentation.
EXPOSITION quality **Score**: 9/10	**Well done**: In English. "The good an the bad slide" is the kind of slides that help get attention. The students gave good answers to the questions. **Can be improved**: They start too flat, not even clarify what the goal is. They should think of something that catches attention, rather than a summary. For instance, the conclusions at the end ask questions, some of which could be emphasised at the beginning. The speakers should do a smoother handover between them, keeping the flow. In the end, the presentation does not look fully integrated; it lacks a narrative. It would be good to have some take-aways at the end. Regarding the style of the presentation, the three speakers should transmit more enthusiasm. Plots are very simple and rough and slides overall could be improved significantly (the first slides are screenshots from R and the rest is very sketchy).

The second project had to do with a better understanding of the exposure of drivers to traffic risks along specified routes, including the proposal of safe alternative routes. For the first assessment, the level of readiness of the project was very high and the students did a good job on the establishment of the main requirements, goals and motivation. They carried out a correct and illustrative exploratory data analysis, preprocessing, modelling and evaluation, as well as a clear exposition of the lessons learnt. The students were happy with their average

TABLE IV
FEEDBACK PROVIDED FOR A MATURE PROJECT ON THE THE ANALYSIS OF TRAFFIC RISKS FOR DRIVERS ALONG SPECIFIED ROUTES.

Item	Feedback
Data VALUE	**Well done**: There are two main areas for value, recommending safe routes and the analysis of he data for policy making. Some interesting findings were discovered or confirmed, such as the influence of age.
Score: 9/10	**Can be improved**: More value could be obtained in other ways, especially if different recommendations are done depending on age or the same route but with speed recommendations, etc.
ALTERNATIVES and innovation	**Well done**: They have compared with RouteWise (safety-based route navigation) and papers.
Score: 8/10	**Can be improved**: Traffic/accident data is analysed by governments and insurance companies. It is difficult to be novel here, but the use of external information (e.g., density, industrial activity, age per area, etc.) could make the analysis different.
TECHNICAL tool integration	**Well done**: The students have integrated ideas from the course. Maps were used at several points.
Score: 8/10	**Can be improved**: More modelling, such as ways of predicting the probability of accident of a route taking into account the conditions, the age of the driver, etc.
Project EFFORT	**Well done**: They work with large databases, an important effort in data preparation and attribute transformation.
Score: 9/10	**Can be improved**: The route recommendation model could have been explored further.
EXPOSITION quality	**Well done**: I really liked the "1.2 Million" slide, as it attracts the attention of the audience to an important problem. They focused on surprising findings and their explanation (such as children driving). Good examples (e.g., roundabouts). Good answers to the questions.
Score: 9/10	**Can be improved**: Too much time devoted on the weather attribute.The plots used for representing slight vs serious are not the best option, as everything is relative (proportion) but we don't see the magnitudes.

score after their first presentation (8.6/10) and decided not to improve and resubmit their work in the final assessment. This was the final grade for the project (before using the co-evaluation adjustments).

VII. RESULTS AND EVALUATION OF THE PBL METHODOLOGY

A. Application of the rubric

Figure 2 shows the average results of the assessments from the course CDA in 2021. The course had a total of 29 students, which formed 10 groups. The red series (*one-shot assessment*) in Figure 2 represents the average performance (with confidence intervals) of the students that were happy with their score in the pre-assessment and decided not to do the resit (8 groups and 25 students in total). The blue series (*pre-assessment*) represents the average performance on the pre-assessment phase for those students that decided to do the resit (2 groups and 4 students in total). Finally, the green series (*resit*) represents the performance of the previous students in the resit after receiving feedback.

Figure 2 shows that for rubric items such as "Project effort", "Technical tool integration" and "Exposition", students did extremely well. Interestingly, results show that below-par students perform very poorly on the rubric item "Alternatives and Innovation", where students should perform a scientific literature search and market research for alternative proposals, also checking how their project is placed with respect to the state of the art. This item significanly improved after the feedback provided. It seems surprising that the ability to look at alternatives is such an important predictor of team performance.

At the individual level, we see in Figure 3 that the project grades do not correlate much (Spearman's rank-order correlation) with those obtained in the average score of the three practical assessments (0.21) and in the two questionnaires (0.05). However, the correlation between the practicals and questionnaires is much higher (0.47). The low correlation may simply come from the fact that the practices and questionnaires are graded individually and the group grade is collective, so it depends on the other members of the group. This may also be understood as an indication that PBL may lead to a lower performance of students on certain (more traditional) course outcomes while still obtaining better performance on those more experimental, pragmatic and real-world tasks.

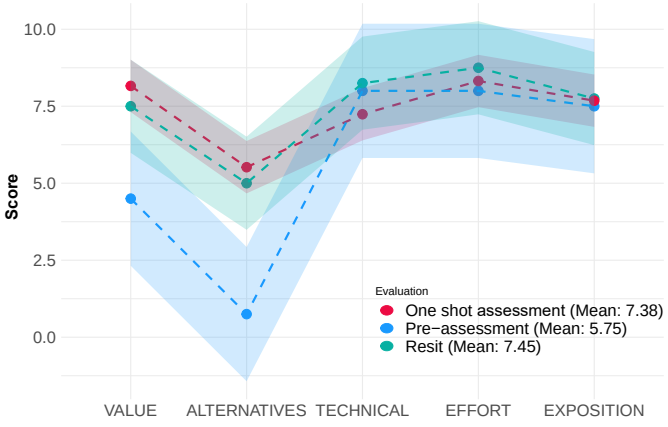

Fig. 2. Students' performance on the different rubric items (for course CDA Fall 2021). Average score summarised by evaluation procedure/phase: one-shot assessment (red), pre-assessment (blue) and final assessment (green).

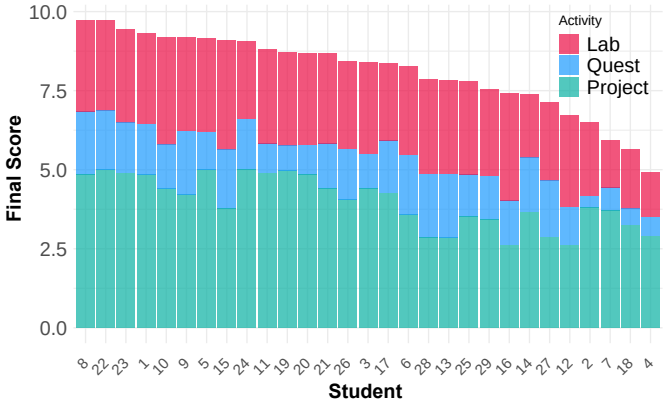

Fig. 3. Students' final scores, broken out by activity.

For its part, as mentioned in section V, the evaluation of the final project through the rubric also allows us to assess the acquisition of transversal competences by means of the items included. In this regard, the final marks obtained in the corresponding items will be used to obtain the values required by the university, which has to be expressed as a rating stated on a scale of 4 values: *Poor* ($0 \leq x \leq 0.3$), *Fair* ($0.3 < x \leq 0.5$), *Good* ($0.5 < x \leq 0.8$) and *Excellent* ($0.8 < x \leq 1.0$). Figure 4 shows the overall results achieved. In general terms, the majority of students have satisfactorily acquired the different transversal competences. This coincides with the the personal perception of the teachers of the course

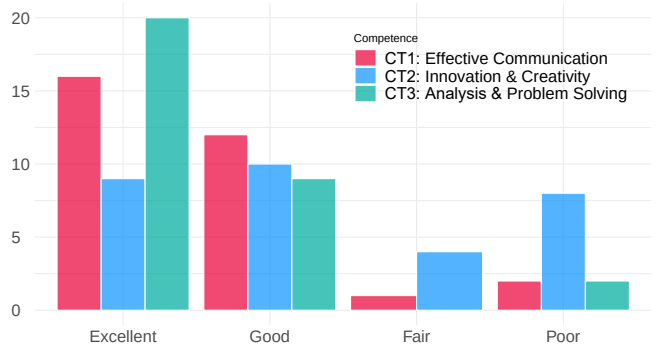

Fig. 4. Competence evaluation through the items in the Rubric in Table I.

in comparison with other courses taught by them where more traditional teaching methodologies are followed. We encounter the greatest problems with the competence of innovation and creativity, where around 25% of the students did not acquire the competence in the expected way. The poor performance of these students in the rubric item "Alternatives and Innovation" (already shown in Figure 2) shows a clear aspect of the course that we can try to improve in the following years (e.g., by fostering brainstorming sessions with teammates, research activities, incorporating further feedback from teammates and teachers, working in class on related issues such as dealing with uncertainty, ambiguity, independence, tenacity, etc.).

B. Course methodology survey

In order to assess whether the teaching methodology has increased student motivation to achieve the competences set out in the data science course, we have developed a questionnaire to be filled in by the students via Google Forms. The questionnaire contains 10 questions, 8 of them based on 5-point likert-type rating scales (from "Strongly disagree" to "Strongly agree") and the remaining 2 being open-ended. Completing the form was optional and we received a total of 20 responses. The details of the questions can be found in Table V.

Questions Q1 to Q3 refer to the motivation of the students with regard to the methodology followed in the course. Questions Q4 to Q8 refer to the objectives and competences of the course. Finally, questions Q9 and Q10 try to bring out critical thinking about the methodology followed in the course.

Figure 5 shows the results of the questions based on rating scales (Q1 to Q8). In the questions referring to the motivation block (Q1 to Q3), we see that 5% of the students marked questions Q2 and Q3 as "Disagree". For its part, 85% of the students consider that the teaching methodology has served to motivate them in learning data science and considered the data science project proposed in the course interesting. Finally, 75% of the students consider that having worked collaboratively in a team has improved not only their participation but also their interest in the subject.

The second block of questions (Q4 to Q8) shows that the students also broadly agree that they have achieved the objectives and competences set by the course. The questions with some disagreement are Q5 and Q6. Question Q5 shows

TABLE V
TEACHING METHODOLOGY EVALUATION QUESTIONNAIRE.

Q	Question	Answer
1	Did the way the course was developed motivate you?	Likert (1-5)
2	Did you find the self-driven data science project proposed in this course interesting?	Likert (1-5)
3	Do you think that your involvement and interest in the course has been increased by working with your classmates in a team?	Likert (1-5)
4	The aim of the final project is for you to understand the role of the data scientist in organisations, identify problems and opportunities and deploy solutions using commonly used tools. Do you think these objectives have been met after the course?	Likert (1-5)
5	After having completed and presented the final project, do you feel more confident to apply what you have learnt in your future work?	Likert (1-5)
6	Do you think that the completion of the final project has allowed you to work on skills such as innovation, creativity and entrepreneurship?	Likert (1-5)
7	Do you think that carrying out the final project has allowed you to work on problem-solving skills in an autonomous way?	Likert (1-5)
8	Do you think that the final project has allowed you to work on effective communication and public speaking skills?	Likert (1-5)
9	Mention positive aspects of this course	Open
10	Mention negative aspects of this course	Open

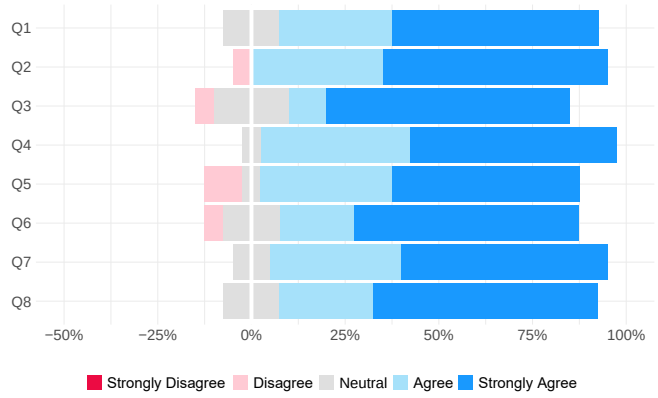

Fig. 5. Summary of results for the survey in Table V. 20 responses processed.

that, on average, there are 10% of students who do not feel confident in applying what they have learnt during the course. This could mean two things: either that their professional area is very far away from data science, or that they still do not see themselves qualified to work as data scientists and need to go deeper into the subject with more advanced courses. The data for Q6 shows that there is a small percentage (5% of students) who claim not to have acquired or improved skills such as innovation, creativity and entrepreneurship. This may be due to a lack of curiosity or inspiration by the student which could also be affected by a lack of understanding of how to properly conceptualise and develop the project idea. Regarding Q7-Q8, no one marked them with values in the disagreement range, which indicates that the competences of problem solving and effective communication have been worked on appropriately during the course. Overall, all the answers indicate a high degree of agreement, where all the above questions have over 75% of positive answers, some of them (e.g., Q4 and Q5), close to 100%. These data show that the vast majority of students consider that both the objectives and the competences set at the beginning have been achieved with the teaching

methodology.

In addition to the Likert-type questions, Q9 and Q10 were used to elicit the positive and negative aspects of the teaching methodology. Below, we summarise the most relevant ideas that we extracted from both questions. Regarding the positive aspects:

- Teamwork and coordination are valued.
- Creativity and problem solving is enhanced.
- Students praise the structure of the course, the challenging and exciting practicals, as well as the real utility of the final project.
- The solid grounding in data science fundamentals is useful.

As for the negative aspects, we highlight the following:

- It was difficult for some students to find interesting ideas and data for the final project.
- Some students lacked time to fully develop some parts of their project.
- Some students considered that the deadlines for practicals and project delivery were too tight.

Overall, the majority of the students have recognised that the methodology followed and the participation and development of projects have pushed them to learn through collaboration, research and better communication, improving also their interpersonal skills.

We usually also receive positive feedback from the students from the university surveys. Apart from some scores, these surveys also provide useful information about the instructors' performance. The survey had a total of 10 questions that aimed to analyse different aspects of the subject, as well as to try to verify the consistency of the answers, in order to avoid random completion. Concretely, with a general average satisfaction 8.5/10, the majority of the students recognised that the instructors show a true commitment and interest in their students, know how to be flexible and attend to individual needs.

VIII. Conclusions

This article has presented the experience, innovations and lessons learned from the application of a PBL strategy to a data science subject. In fact, this work provided a set of innovative didactic resources, assessment tools and knowledge that could be used to organise data science courses in general. We have seen how the absence of rigid schemes and the exploratory character of data science, as well as the range of proactive, curious and inquisitive skills that data scientists should have, makes the adoption of PBL to be more advantageous, but also more challenging, precisely because specifications and stages are not clear as in other engineering projects. Still, the obtained results follow the same line in terms of the effectiveness and relevance of PBL methodologies in other computing and engineering disciplines (see, e.g., [4], [5], [8]).

The aim for the students is to learn how to develop projects themselves considering also the relevant role of data value and innovation. By applying our rubric-based assessments, we also helped to ensure that the discussions and feedback with students are more effective, also making the monitoring and evaluation more systematic. The teaching methodology has been evaluated through a questionnaire filled in by the students and the analysis of the grades obtained in the activity. Both elements show that the objectives and the competences necessary for working as data scientists have been achieved. Also, according to the answers to the questionnaire, it seems that the methodology followed has increased the motivation of the students, capturing their interest and attention. This perception is also shared by the teachers, based on their experience in teaching other courses on similar subjects which do not follow a PBL methodology.

While the application of this teaching methodology did generate positive outcomes for project-based data science courses, more work needs to be done to further compare the results of this PBL methodology with other potential pedagogical approaches. For example, one might compare case studies or Kaggle-like competitions[2] with this project-focused course. Also, this teaching methodology can still be refined and improved. Based on student and university feedback, specific refinements may include reducing, but not eliminating, the amount of self-learning that needs to occur for students to be successful in the class as well as the number of laboratory practicals to be carried out (considering the perceived effort in addition to the final project). Also, further activities intended to enhance the innovation skills of the students will have to be incorporated in the course, as it is the part where students tend to perform worst according to the assessment rubrics. On the other hand, student commitment for improving and re-submitting the final project can be increased further through the possibility of allowing them to participate in a closing event, workshop or project fair where students may submit and present their work to other students, industrial partners, and interested third parties.

As future work, we also want to consider and encourage the role of automation in these projects. As in other areas using AI [41], it is becoming more common that students use sophisticated tools that accelerate their process, from data-wrangling to Auto-ML [42]. How to balance the 'project effort' entry in the rubric with a smarter use of tools that reduce data scientists' effort require a very clear statement about how to declare the use of these tools and how they are positively graded.

Acknowledgment

We thank the anonymous reviewers for their comments. This work was funded by valgrAI, the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI, the Future of Life Institute, FLI, under grant RFP2-152, the EU (FEDER) and Spanish grant RTI2018-094403-B-C32 funded by MCIN/AEI/10.13039/501100011033 and by CIPROM/2022/6 funded by Generalitat Valenciana, EU's Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), US DARPA HR00112120007 (RECoG-AI) and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe".

[2]https://www.kaggle.com/

REFERENCES

[1] D. Nunan, C. Candlin, and H. Widdowson, Syllabus design. Oxford University Press Oxford, 1988, vol. 55.

[2] J. D. Bigelow, "Using problem-based learning to develop skills in solving unstructured problems," Journal of Management Education, vol. 28, no. 5, pp. 591–609, 2004.

[3] M. A. Almulla, "The effectiveness of the project-based learning (pbl) approach as a way to engage students in learning," Sage Open, vol. 10, no. 3, 2020.

[4] R. Pucher and M. Lehner, "Project based learning in computer science– a review of more than 500 projects," Procedia-Social and Behavioral Sciences, vol. 29, pp. 1561–1566, 2011.

[5] M. L. Fioravanti, B. Sena, L. N. Paschoal, L. R. Silva, A. P. Allian, E. Y. Nakagawa, S. R. Souza, S. Isotani, and E. F. Barbosa, "Integrating project based learning and project management for software engineering teaching: An experience report," in SIGCSE, 2018, pp. 806–811.

[6] J. A. Macías, "Enhancing project-based learning in software engineering lab teaching through an e-portfolio approach," IEEE Transactions on Education, vol. 55, no. 4, pp. 502–507, 2012.

[7] M. Souza, R. Moreira, and E. Figueiredo, "Students perception on the use of project-based learning in software engineering education," in 33th Brazilian Symp. on Software Engineering, 2019, pp. 537–546.

[8] J. W. McManus and P. J. Costello, "Project based learning in computer science: a student and research advisor's perspective," Journal of Computing Sciences in Colleges, vol. 34, no. 3, pp. 38–46, 2019.

[9] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. H. Orallo, M. Kull, N. Lachiche, M. J. R. Quintana, and P. A. Flach, "CRISP-DM twenty years later: From data mining processes to data science trajectories," IEEE Trans. on Knowledge and Data Engineering, 2019.

[10] J. Hernández-Orallo, C. Ferri, and M. Ramírez-quintana, Introduction to Data Mining. Pearson, 2004.

[11] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 step-by-step data mining guide," 2000.

[12] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, P. Flach, J. Hernández-Orallo, M. Kull, N. Lachiche, and M. J. Ramírez-Quintana, "CASP-DM: context aware standard process for data mining," arXiv preprint arXiv:1709.09003, 2017.

[13] M. O'Neil, "As data proliferate, so do data-related graduate programs," The Chronicle of Higher Education, vol. 60, p. 1, 2014.

[14] R. E. Anderson, M. D. Ernst, R. Ordóñez, P. Pham, and B. Tribelhorn, "A data programming cs1 course," in 46th ACM Technical Symp. on Computer Science Education, 2015, pp. 150–155.

[15] Y. Demchenko, A. Belloum, C. de Laat, C. Loomis, T. Wiktorski, and E. Spekschoor, "Customisable data science educational environment: From competences management and curriculum design to virtual labs on-demand," in 2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom). IEEE, 2017, pp. 363–368.

[16] J. W. Johnson, "Scaling up: Introducing undergraduates to data science early in their college careers," Journal of Computing Sciences in Colleges, vol. 33, no. 6, pp. 76–85, 2018.

[17] Y. M. Kim, "The effects of pbl-based data science education program using app inventor on elementary students' computational thinking and creativity improvement." Ilkogretim Online, vol. 20, no. 1, 2021.

[18] C. Dichev, D. Dicheva, L. Cassel, D. Goelman, and M. Posner, "Preparing all students for the data-driven world," in Symposium on Computing at Minority Institutions, ADMI, vol. 346, 2016.

[19] C. Vera, J. Félez, J. Antonio Cobos, M. J. Sánchez-Naranjo, and G. Pinto, "Experiences in education innovation: developing tools in support of active learning," European Journal of Engineering Education, vol. 31, no. 2, pp. 227–236, 2006.

[20] D. A. Asamoah, D. Doran, and S. Schiller, "Interdisciplinarity in data science pedagogy: a foundational design," Journal of Computer Information Systems, vol. 60, no. 4, pp. 370–377, 2020.

[21] B. Cassel and H. Topi, "Strengthening data science education through collaboration," in WS on data science education, vol. 7, 2015, p. 27.

[22] Q. Cheng, F. Lopez, and A. Hadjixenofontos, "Integrating introductory data science into computer and information literacy through collaborative project-based learning," in 2019 IEEE Frontiers in Education Conference (FIE). IEEE, 2019, pp. 1–5.

[23] D. White, "A project-based approach to statistics and data science," Primus, vol. 29, no. 9, pp. 997–1038, 2019.

[24] L. Philip and W. K. Li, "Project-based learning via competition for data science students," Harvard Data Science Review, 2021.

[25] N. Postman and C. Weingartner, "Teaching as a subversive activity. delacorte press," New York, 1969.

[26] H. S. Barrows, R. M. Tamblyn et al., Problem-based learning: An approach to medical education. Springer, 1980, vol. 1.

[27] J. Kay, M. Barg, A. Fekete, T. Greening, O. Hollands, J. H. Kingston, and K. Crawford, "Problem-based learning for foundation computer science courses," CSE, vol. 10, no. 2, pp. 109–128, 2000.

[28] P. Dart, L. Johnston, and C. Schmidt, "Enhancing project-based learning: Variations on mentoring," in Proceedings of 1996 Australian Software Engineering Conference. IEEE, 1996, pp. 112–117.

[29] W. J. Pluta, B. F. Richards, and A. Mutnick, "Pbl and beyond: Trends in collaborative learning," Teaching and learning in medicine, vol. 25, no. sup1, pp. S9–S16, 2013.

[30] M. H. Baturay and O. F. Bay, "The effects of PBL on the classroom community perceptions and achievement of web-based education students," Computers & Education, vol. 55(1), pp. 43–52, 2010.

[31] Y. Woo and T. C. Reeves, "Meaningful interaction in web-based learning: A social constructivist interpretation," The Internet and higher education, vol. 10, no. 1, pp. 15–25, 2007.

[32] C. Hayashi, "What is data science? fundamental concepts and a heuristic example," in Data science, classification, and related methods. Springer, 1998, pp. 40–51.

[33] D. Holtz, "8 Skills You Need to Be a Data Scientist," https://blog.udacity.com/2014/11/data-science-job-skills.html, 2014.

[34] V. Dhar, "Data science and prediction," Communications of the ACM, vol. 56, no. 12, pp. 64–73, 2013.

[35] M. Loukides, What Is Data Science? O'Reilly Media, Apr. 2011.

[36] E. Commission, "European e-Competence Framework," 2016. [Online]. Available: http://www.ecompetences.eu/

[37] M. J. Sá and S. Serpa, "Transversal competences: Their importance and learning processes by higher education students," Education Sciences, vol. 8, no. 3, p. 126, 2018.

[38] R. Pucher, A. Mense, and H. Wahl, "Intrinsic motivation of students in project based learning in South Africa," SAIEE, vol. 94, pp. 7–14, 2003.

[39] F. Provost and T. Fawcett, Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013.

[40] H. Harris, S. Murphy, and M. Vaisman, Analyzing the analyzers: An introspective survey of data scientists and their work. " O'Reilly Media, Inc.", 2013.

[41] F. Martínez-Plumed, S. Tolan, A. Pesole, J. Hernández-Orallo, E. Fernández-Macías, and E. Gómez, "Does AI qualify for the job? a bidirectional model mapping labour and ai intensities," in AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 94–100.

[42] T. D. Bie, L. D. Raedt, J. Hernández-Orallo, H. H. Hoos, P. Smyth, and C. K. I. Williams, "Automating data science: Prospects and challenges," Communications of the ACM, vol. 65, no. 2, pp. 76–87, 2022.