Challenge-based learning in Computational Biology and Data Science

Emilio Serrano*, Martin Molina, Daniel Manrique**, Javier Bajo
{emilioserra,mmolina,dmanrique,jbajo}@fi.upm.es

Department of Artificial Intelligence, Universidad Politécnica de Madrid, Spain

Abstract. Data Science is an interdisciplinary field devoted to extract knowledge from large amounts of data. There is a great variety of programs that address the teaching of this field with a growing demand of professionals. However, data science pedagogy tends to emphasize general aspects of data and the use of tools instead of the its scientific dimension. This position paper describes an ongoing educational innovation project for the use of the Challenge-based Learning approach to teach and learn Data Science. In this approach, students work on solving complex and real world problems while the learning is obtained by iterating through three main phases: engage, investigate, and act.

Keywords: Challenge-based learning, active learning, experiential learning, project based learning, data science, computational biology.

1 Introduction

Data science (DS) is an interdisciplinary field devoted to identify patterns and extract knowledge by mining large amounts of structured and unstructured data. Among others, DS includes: machine learning, data processing, statistical research, and their related methods. This science has become a revolution that has changed our manner of doing business, health, politics, education and innovation [11]. Scientific breakthroughs will be increasingly assisted by advanced computing capabilities and DS methods that help researchers manipulate and explore massive datasets [9].

Challenge-based learning (CBL) is a new learning approach created by Apple Inc. in collaboration with teachers and leaders in the education community. CBL is “an engaging, multidisciplinary approach that starts with standards-based content and lets students leverage the technology they use in their daily lives to solve complex, real-world problems” [5]. In CBL, students work with other students, their teachers, and experts in their communities and around the world to develop deeper knowledge of the subjects they are studying.

Data science is in a privileged position with respect to other branches of knowledge to articulate learning through experiences and challenges [18]. The

* ORCID ID: 0000-0001-7587-0703

** ORCID ID: 0000-0002-0792-4156

Kaggle platform [2] periodically releases a series of competitions on real problems such as “Predicting a Biological Response” [4]; which offered 20,000$ to the best predictive model that linked a biological response of molecules to their chemical properties. These public competitions have the potential to involve actively the student in a real, significant, and related problematic situation; including a framework for the implementation of a solution to the challenge.

This position paper presents an ongoing educational innovation project focusing on using the CBL approach in a DS course, as part of a Computational Biology master degree at the Technical University of Madrid (UPM). Students will work on challenges at the level of a Kaggle competition with special preference for active and multidisciplinary problems. Based on the 2016 update for the CBL framework proposed by Apple Inc. [12], students will learn by following three main phases: engage, investigate, and act.

The paper outline is as follows: after describing the background of the presented innovation project in section 2, the project details are given in section 3. These include the scope and students' profile, the project goals, the timeline and educational resources, the evaluation, resulting products, and diffusion plan. Section 4 explains the expected contribution to the improvement of learning quality. Finally, section 5 concludes and presents future lines of research and work.

2 Background

The great diversity of applications and the growing demand of experts in the DS field has made courses, books and manuals in DS proliferate [18]. The standard pedagogical method that we can appreciate in these courses consists of four steps:

1. The explanation of the different machine learning branches (supervised, unsupervised, and by reinforcement).
2. The detail of some learning paradigms under some of these branches; such as decision trees or artificial neural networks.
3. The illustration of these paradigms using toy datasets such as Weather or Iris [21].
4. Assignments with a straightforward application of the ideas previously exposed using some DS framework such as Weka [7] or Caret [8].

We executed an educational innovation project last year, where the limitations of this standard pedagogical approach were revealed [17]. Instead, an experiential learning (EL) method was successfully adopted in a Deep Learning course, included in the Master in Data Science (EIT Digital Master School), offered at UPM.

EL brings real life experiences into the classroom which must be integrated with the goals and objectives of the discipline theory [13]. The students reflecting on their product is a fundamental part of EL [20]. Different learning approaches based on the Kolb cycle [10] were proposed, applied, and evaluated in the deep learning course within the frame of the educational innovation project. According to this cycle, effective learning involves: having a concrete experience, observation of and reflection on that experience, the formation of abstract concepts (analysis) and generalizations (conclusions), and testing them by active experimentation, resulting in new experiences (iterations in the cycle).

Some of the results of this previous project [17] were presented in a position paper for an international conference [18], a Spanish conference paper [19], and software tool to complement *JupyterHub for Teaching*[1].

3 Project details

This section explains the ongoing project details [16] to allow interested professors to extrapolate our case to their specific environment. In this new project, the CBL approach is studied in a new course with different students' profiles.

3.1 Scope and students' profile

The project will be developed with students attending the master in Computational Biology, offered by the Computer Science School at the UPM. More specifically, in the module of "Knowledge representation and acquisition".

The profile of these Computational Biology students is multidisciplinary, belonging part of them to the world of biology and part of them to branches of information technology. Data science and CBL provide an exceptional framework to establish synergies between these two main profiles, e.g. applying computer science techniques to biology or vice-versa. In this vein, one of the requisites of the challenges will be to include both profiles in each work group.

Students will also have the opportunity to apply the knowledge acquired from other master's modules such as "Statistical Analysis and Data Visualization" or "Machine Learning" to real world problems. The results of this project will be directly applicable to other modules, whether of this master degree or of the master in Data Science (EIT Digital Master School)[2], and also in courses of the Bachelor Degree in Computer Engineering[3] such as "Data Mining".

3.2 Goals

The goals of the presented project are the following:

- G1. Development of methods for CBL in Data Science and Computational Biology. This goal includes the instantiation of methodologies and general frameworks of CBL to the specific field to be treated. Among these frameworks, we can point out: the 2016 update for the CBL framework proposed by Apple Inc [12]; and the "Challenge Based Learning" report carried out

[1] `https://jupyterhub-deploy-teaching.readthedocs.io`

[2] `http://www.fi.upm.es/?id=masterdatascience`

[3] `https://www.fi.upm.es/?id=gradoingenieriainformatica&idioma=english`

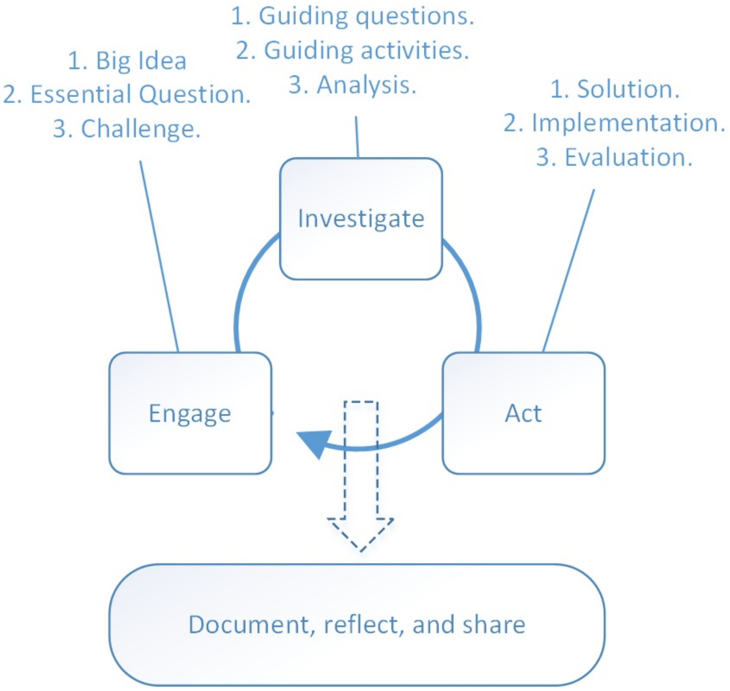


Fig. 1: CBL framework proposed by Apple Inc.

by the Monterrey Institute of Technology and Higher Education in 2016 [6]. The phases of the Apple Inc framework are depicted in Fig. 1.

- G2. Study of specific challenges in the field of Data Science and Computational Biology. Application of the explored, extended, and instantiated methods in O1 to Data Science. This goal addresses the selection of concrete challenges and for a specific student profile. The students will work on a challenge at the Kaggle competition level with special preference for active and multidisciplinary problems. Examples of past competitions that fit the students' profile are "Predicting a Biological Response", "Merck Molecular Activity Challenge", "Shelter Animal Outcomes", "Leaf Classification", or "Zoo Animal Classification".
- G3. Integration and documentation of tools for the support of CBL in the course of "Representation and Acquisition of Knowledge". This goal includes the analysis and documentation of tools available to support the CBL in this specific module. Although Kaggle has a working environment, Kaggle Kernels[4], this framework will be combined with other tools, preferably free and open source, that meet the needs of the CBL. Among others and according to the 2016 update for the ABR framework proposed by Apple Inc. [12],

[4] https://www.kaggle.com/kernels

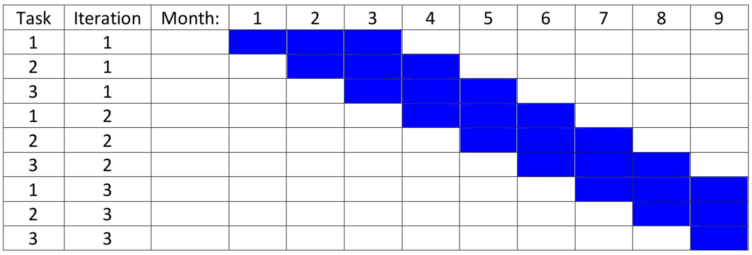
students will need: a calendar, space for collaboration, and storage of documents. Project management tools such as Trello[5] and Asana[6] as well as free alternatives will be considered.

3.3 Timeline and educational resources

The following three major tasks will be addressed with a clear correspondence to the three goals explained above:

- T1. Developing of methods for CBL in Data Science and Computational Biology.
- T2. Analyzing and selecting specific challenges in the field of Data Science and Computational Biology.
- T3. Integrating and documenting tools for the support of CBL in the Representation and Acquisition of Knowledge.

The project kicks-off on February 15, 2018 and an ends on November 15, 2018, giving 9 months of project numbered from 1 to 9. In an iterative and incremental approach such as the Scrum methodology [15, 14], the following timeline is proposed in Fig. 2.



Task	Iteration	Month:	1	2	3	4	5	6	7	8	9
1	1		🟦	🟦	🟦						
2	1			🟦	🟦	🟦					
3	1				🟦	🟦	🟦				
1	2					🟦	🟦	🟦			
2	2						🟦	🟦	🟦		
3	2							🟦	🟦	🟦	
1	3								🟦	🟦	🟦
2	3									🟦	🟦
3	3										🟦

Fig. 2: Timeline of the project with an iterative and incremental approach.

As shown in Fig. 2, three iterations are considered for each task. This allows the tasks results to feed back to previous tasks. Moreover, there is an overlap into the different tasks as expected in an iterative and incremental approach.

The following educational resources will be used:

- Scientific repositories available in the UPM as ScienceDirect[7].
- The Institutional Teaching Platform of the UPM (Moodle)[8].
- Data repositories and contests websites in Data Science such as Kaggle.
- Resources of the Department of Artificial Intelligence as web servers.

[5] https://trello.com/

[6] https://asana.com

[7] www.sciencedirect.com

[8] moodle.upm.es

3.4 Evaluation

The evaluation of the project will be carried out through the generation of an *e-portfolio* by the students attending the module under study. The e-portfolio is a digital collection of evidence which includes: demonstrations, resources, and achievements obtained by students. According to the CBL framework proposed by Apple Inc [12], the following sections will be evaluated:

- Report of the great ideas to investigate.
- The proposal of the challenge, the essential question to answer and the motivation about the significance of the challenge.
- Guiding issues, questions that will guide the search for a solution.
- Learning plan and schedule.
- Research report, in Jupyter IPython notebook format[9] or alternative to ensure the reproducibility and repeatability of the results achieved.
- Proposed solution, presentation including prototypes, concepts, and expert feedback.
- Implementation and evaluation plans.
- Evaluation results.
- Final presentations.
- Journals with personal and group experience.
- Final reflections on what was learned.

3.5 Resulting products

This section describes the tangible products resulting from the project (methodological guides, reports, educational resources, etcetera) with a description of their potential for internal and external transfer. The following deliverables will be elaborated:

- D1. Report on methods for CBL in Computational Biology and Data Science.
- D2. Report on the analysis and selection of appropriate challenges for the learning of Computational Biology and Data Science.
- D3. Manual of tools for the support to the CBL in the Representation and Acquisition of Knowledge.
- D4. Report on the evaluation of results based on e-portfolios created by students.
- D5. Journal and conference papers for the dissemination of results.

As described in section 3.1 these deliverables have an internal transfer in the UPM to other modules of the master for which it is proposed, other masters, and other degrees. These products can also be a competitive advantage in the organization of massive open online courses (MOOCs) by presenting a pedagogical prescriptions.

[9] `jupyter.org`

3.6 Diffusion plan

The main diffusion materials generated in the project will be the deliverables 3, 4 and 5 explained in section 3.5. There will also be contemplated:

- the construction of a website that collects all the deliverables,
- news for diffusion at the UPM,
- microblogging posts (Twitter) in the department and the school,
- radio interviews to disseminate educational innovation.

4 Contribution to the improvement of teaching quality

Thanks to the application of the CBL approach, a significant improvement of learning and teaching quality is expected. Among others, this improvement will be reflected in:

1. A deeper understanding of Data Science for Computational Biology, allowing to diagnose and analyze problems before proposing solutions.
2. A greater commitment to involve the student both in the definition of the Data Science problem to be addressed and in the solution that will be developed to solve it.
3. Development of skills to investigate, create models, materialize them, and work collaboratively and multidisciplinary.
4. A closer approach to the reality of their profession, establishing relationships with specialists in the Kaggle platform that contribute to their professional growth.
5. Strengthening the connection between what they learn in the Master's and what they perceive in the professional world.
6. Development of high-level communication skills, through the use of social tools such as Kaggle forums and media production techniques, to create and share the solutions developed by them.

Moreover, this teaching innovation project [16] will be aligned with the results obtained from our previous project using EL in Data Science [17, 18]. Therefore, the same benefits obtained in the Deep Learning course are expected in the "Knowledge representation and acquisition" module. Among others, the student will:

1. learn to select relevant information about how learning paradigms work and the information they offer, instead of considering them as black boxes where the model built has no relevance and only quality metrics are studied;
2. learn to study the details and data of the concrete problem and to obtain good understanding of the data;
3. perceive the iterative nature of DS, by building different prediction models considering the data and results from previous models;
4. and, research on new methods and their extension or variation for new and challenging problems, instead of just applying well-known solutions to well-known problems.

5 Conclusion and future works

This position paper has presented an ongoing educational research project to use the challenge-based learning approach for DS in the context of a master in Computational Biology. The paper has revised the background, including a number of shortcomings repeated in current DS courses, and the results obtained in a previous project, based on considering experiential learning for DS. The ongoing project details have been presented: the scope and students' profiles, goals, timeline, teaching resources, evaluation, resulting products, and diffusion plan. The project details allow interested professors to extrapolate our case to their specific environment and audience.

The contribution to the improvement of the teaching quality of the project has also been explored, highlighting a deeper understanding of Data Science for Computational Biology. This allows students to diagnose and analyze problems before proposing solutions. DS is not only about data and tools to manage them as classic DS courses may suggest. DS is more about "science" and the scientific questions we can answer with data. Therefore, a major advantage in using CBL for DS is that teachers do not present the answers before students ask the scientific questions by themselves.

Our main future works include to create a survey on the acceptance and quality of challenges proposed, to study new theoretical frameworks for applying CBL in DS, and the exploration (or implementation) of software tools to develop e-portfolios. Moreover, two students have been hired to assist in the search and development of specific challenges for Computational Biology.

Acknowledgments

This research work is supported by the Universidad Politécnica de Madrid under the education innovation project "Aprendizaje basado en retos para la Biología Computacional y la Ciencia de Datos", code IE1718.1003; and by the the Spanish Ministry of Economy, Indystry and Competitiveness under the R&D project Datos 4.0: Retos y soluciones (TIN2016-78011-C4-4-R, AEI/FEDER, UE).

References

1. Data Science Specialization, Johns Hopkins University. https://www.coursera.org/specializations/jhu-data-science. Accessed: March of 2018.
2. Kaggle: Academic Machine Learning Competitions. https://inclass.kaggle.com/. Accessed: May of 2017.
3. Machine Learning MOOC, Stanford University. https://www.coursera.org/learn/machine-learning. Accessed: March of 2018.
4. Predicting a Biological Response. https://www.kaggle.com/c/bioresponse. Accessed: March of 2018.
5. Challenge Based Learning. A Classroom Guide. goo.gl/vAwsg8, 2011. Accessed: March of 2018.

6. Edu Trends. Aprendizaje Basado en Retos. `goo.gl/dA3ux8`, 2016. Accessed: March of 2018.
7. E. Frank, M. A. Hall, G. Holmes, R. Kirkby, and B. Pfahringer. Weka - a machine learning workbench for data mining. In O. Maimon and L. Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 1305–1314. Springer, 2005.
8. M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, and A. Engelhardt. *caret: Classification and Regression Training*, 2012. R package version 5.15-044.
9. A. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research, 2009.
10. D. A. Kolb. *Experiential Learning: Experience as the Source of Learning and Development*. Prentice Hall, 1 edition, Oct. 1983.
11. V. Mayer-Schonberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, 2013.
12. M. Nichols, K. Cator, and M. Torres. *Challenge Based Learner User Guide*. Redwood City, CA: Digital Promise, 2016.
13. Qualters and C. Wehlburg. *Experiential Education: Making the Most of Learning Outside the Classroom: New Directions for Teaching and Learning, Number 124*. J-B TL Single Issue Teaching and Learning. Wiley, 2010.
14. A. R. Santos, A. Sales, P. Fernandes, and M. Nichols. Combining challenge-based learning and scrum framework for mobile application development. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '15, pages 189–194, New York, NY, USA, 2015. ACM.
15. K. Schwaber and M. Beedle. *Agile Software Development with Scrum*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2001.
16. E. Serrano, D. Manrique, J. Bajo, and M. Molina. Aprendizaje basado en retos para la Biología Computacional y la Ciencia de Datos. `https://goo.gl/d7ZwZt`. Accessed: March of 2018.
17. E. Serrano, M. Molina, D. Manrique, and L. Baumela. Métodos, experiencias y herramientas para el aprendizaje experiencial de la Ciencia de Datos. `https://goo.gl/Yy7XeT`. Accessed: March of 2018.
18. E. Serrano, M. Molina, D. Manrique, and L. Baumela. Experiential learning in data science: From the dataset repository to the platform of experiences. In C. Analide and P. Kim, editors, *Intelligent Environments 2017 - Workshop Proceedings of the 13th International Conference on Intelligent Environments, Seoul, Korea, August 2017*, volume 22 of *Ambient Intelligence and Smart Environments*, pages 122–130. IOS Press, 2017.
19. E. Serrano, M. Molina, D. Manrique, L. Baumela, and D. Zanardini. Aprendizaje experiencial en ciencia de datos: satisfacción de los estudiantes para tres modelos de enseñanza y aprendizaje [Experiential learning in data science: student satisfaction for three models of teaching and learning]. 2017.
20. M. Silberman. *The Handbook of Experiential Learning*. Wiley, 2007.
21. I. H. Witten, E. Frank, and M. A. Hall. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, Burlington, MA, 2011.