Tailored Data Science Education using Gamification

Kim Hee, Roberto V. Zicari, Karsten Tolle
Frankfurt Big Data Laboratory
Goethe University Frankfurt
Frankfurt, Germany
{hkim, zicari, tolle}@dbis.cs.uni-frankfurt.de

Andrea Manieri
Ingegneria Informatica spa R&D lab.
Engineering
Rome, Italy
manieri@eng.it

***Abstract*—Interest to become a data scientist or related professions in data science domain is rapidly growing. To meet such a demand, we propose a novel educational service that aims to provide tailored learning paths for data science. Our target user is one who aims to be an expert in data science. Our approach is to analyze the background of the practitioner and match the learning units. A critical feature is that we use gamification to reinforce the practitioner engagement. We believe that our work provides a practical guideline for those who want to learn data science.**

Keywords*— *data science education; tailored learning path; learning goal; learning unit, EDISON competence framework; knowledge area; data crowdsourcing; gamification

I. Introduction

Data Science (DS) is an emerging domain with potential impacts on nearly every aspect of industries and research institutes. According to the report by McKinsey Global Institute [1], "by 2018 the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge". Accordingly, interest to become a data scientist or join a related profession in data science is rapidly growing. In order to meet such high demands from educators, trainers, employers, and research institutes, we propose a fully automated web service that provides a data science custom made training program in a specific order.

However, the implementation of the "tailored" learning paths versus a "random" practitioner is not easy. To foresee this over his learning paths to reach his learning goal due to the three main challenges:

- Each practitioner possesses a different set of skills and demonstrates a different level of proficiency. Furthermore, each practitioner has a unique learning goal.
- There is no generic information on how well various learning units cover data science competences. A common interchange format or information of how the learning units are actually constructed is not available.
- The different existing learning units in Data science might overlap and depend on each other. Practitioners, particularly beginners, may face bottleneck issues due to the complexity of the topics to learn.

We have carefully considered each challenge in order to achieve our service objective. This work is performed as a part of the EDISON project [2], in particular the definition of data science competence framework. EDISON is a two-years EU-funded project that strives to create a foundation for establishing a new profession of Data Scientist for European research and industry.

The first challenge is related to the practitioner. A practitioner is requested to assess their proficiency at 34 competences in data science from 5 mastery levels where "1" is less relevant and "5" is highly relevant. Then a learning goal should be declared by the practitioner. The learning goal is a dedicated profession that is described by a set of competence scores.

The second challenge is related to the different existing training units. Accommodating them in a generic set of learning objectives is an essential part of providing the customized learning paths. We attempted to convert each learning unit into a set of data science competences using data crowdsourcing [3], [4]. Crowdsourcing refers to solving large problems by involving human workers that solve component of sub-problems or tasks [3], while data crowdsourcing particularly focuses on enriching and expanding the initial data set. We use the data crowdsourcing as a data acquisition tool.

Finally, the last challenge is about designing the training program. How can we motivate a practitioner to learn a challenging topic with strong engagement? This is a common instructional objective since an engaged practitioner has a higher degree of attention, curiosity, and passion to meet one's learning objective. We facilitate a gamification [5]–[7] on top of the training program. Gamification is an emerging trend in education to reinforce practitioner's engagement that helps to keep one motivated and engaged through the learning.

The rest of the paper is organized as follows: In Section II we propose a systematic approaching to identify a random practitioner in three steps; in Section III we identify learning units by facilitating a crowdsourcing framework that demands

2330-2186/16 $31.00 © 2016 IEEE
DOI 10.1109/CloudCom.2016.105

contributions of numerous participants instead of a rule-based approach; in Section III.C we provide a game-like learning with five gamification elements. Finally, in Section V, we conclude the paper.

II. Identify a Practitioner

A. Learning Goal

Learning goal in this paper is one of nineteen data science professions grouped into five profession families, which are defined according to the EDISON deliverable 2.1 [2]. We have developed a dynamic HTML page that represents profession families and professions in two levels of hierarchy. A practitioner is requested to select one of the professions that a practitioner wants to be in future after the learning.

Note that the learning goal is not the same as a *learning objective*. A learning goal describes in broad terms what the learners will be able to do upon completion of the path, whereas a learning objective describes, in specific and measurable terms, specific elements that learners will have mastered upon completion of one or more courses [8]. TABLE I provides a mapping between learning goal (profession) and learning objective (competence). It represents relevance of each competence area to the data science profession where "1" is less relevant and "5" is highly relevant.

B. Competences

Understanding a practitioner is an initial task for the tailored service. The practitioner is asked to answer a survey covering the different data science competences. We adopt a definition of data science competences and skills defined in EDISON Deliverable D2.1[2]. These competences are categorized into the following five competence areas that are interchangeable with five knowledge areas:

- *Data Analytics (DSDA)*. Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations.
- *Data Management (DSDM)*. Develop and implement a data management strategy for data collection, storage, preservation, and availability for further processing.
- *Data Science Engineering (DSEN)*. Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis, and management.
- *Scientific and Research Methods (DSRM)*. Create new understandings and capabilities by using the scientific method (hypothesis, test/artifact, and evaluation) or similar engineering methods to discover new approaches to creating new knowledge and achieve research or organizational goals.
- *Data Science Domain Knowledge (DSDK)*. Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain-specific data types and presentations, data and process models, organizational roles and relations.

TABLE I. Mapping professions to the data science competence areas - Data Analytics (DSDA), Data Management (DSDM), Engineering (DSEN), Scientific and Research Methods (DSRM) and Domain Knowledge (DSDK) - including their relevance rates from 1 to 5 [1]

Profession	**Competence/Knowledge Area**				
	DS DA	*DS DM*	*DS EN*	*DS RM*	*DS DK*
Data Science Manager	3	4	3	3	2
Data Science Infrastructure Manager	2	4	4	2	2
Research Infrastructure Manager	2	4	4	3	2
Data Scientist	5	3	4	5	3
Data Science Researcher	4	3	2	5	4
Data Science Architect	4	3	5	3	3
Data Science Programmer/Engineer	4	2	5	3	4
Data Analyst	5	3	3	3	4
Business Analyst	5	3	3	4	5
Data Stewards	3	5	3	3	3
Digital data curator	1	5	2	2	3
Digital Librarians	2	5	2	2	3
Data Archivists	1	5	1	1	3
Large scale database designer	2	4	4	3	3
Large scale database admin	2	4	3	2	3
Scientific database administrator	2	4	3	2	3
Big Data facilities Operator	1	4	4	2	3
Large scale data storage operator	1	4	3	1	1
Scientific database operator	1	4	3	2	3

The practitioner is supposed to choose one's proficiency of each competence from five mastery levels. The level of knowledge achieved through stages of education. Data science (DS) level 3 is considered to be achieved through a bachelor degree, DS level 4 through a master's degree and DS level 5 through a Ph.D. degree. DS levels 1 to 5 are interchangeable with the European e-Competence Framework (e-CF) [9], [10] and compatible to the European Qualifications Framework (EQF) [11]. The mapping and description are shown in TABLE II.

In addition, the current most similar profession is identified by applying similarity function. *Cosine similarity*, one of the most used *similarity function*, is applied to implement this feature. It is sensitive for recognizing whether two data samples are collinear rather than the absolute magnitude of differences. It quantifies the similarity between two numerical inputs from the Section II.A and II.B.

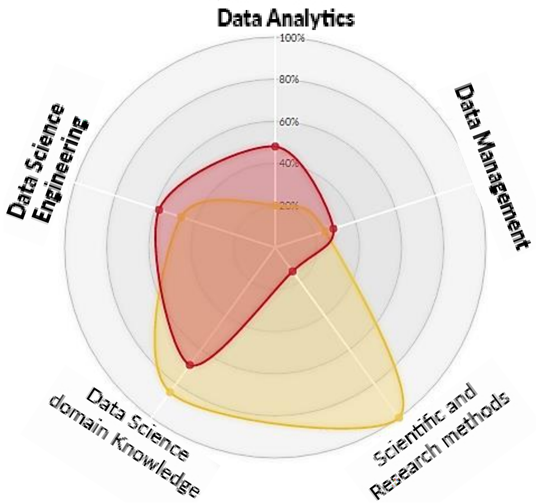
$$\cos(\theta) = \sum_{i=1}^{N} X_i \times Y_i \Bigg/ \sqrt{\sum_{i=1}^{N} {X_i}^2} \times \sqrt{\sum_{i=1}^{N} {Y_i}^2}$$

- X, Y = data points
- Xi, Yi = attributes of each data point to be compared
- N = number of attributes (i.e., dimensionality of X)

TABLE II. Data science mastery levels are interchangeable to the European Qualifications Framework (EQF) and the European e-Competence Framework (e-CF) [1]

DS Level	**EQF Level**	**e-CF Level**	**e-CF Level Description**
5	8	e-5	Principal
4	7	e-4	Lead Professional/Senior Manager
3	6	e-3	Senior Professional/Manager
2	5	e-2	Professional
	4		
1	3	e-1	Associate

Fig 1. Example competence benchmark result. Red polygon indicates the chosen profession and yellow polygon indicates the practitioner. Deficient competences are shown in a bold typeface.



C. Competence Benchmark

A competence benchmark uses two inputs, generated by the participant in the previous step, which is the self-assessment of the data science competences and the selection of the learning goal. It generates a data science competences comparison between the practitioner and the chosen profession. The result of the competence comparison is depicted as two polygons in a spider web as shown in Fig 1. The deficient competences are indicated with a graphical notation in the sake of user-friendly web service. This user-friendly indicator allows the practitioner to identify the proficiency of their competences and which of their competences is sufficient or deficient compared to the chosen profession. For instance, one polygon in the yellow represents proficiency of practitioner's competences; the other in the red indicates the competences of the chosen profession.

III. Identify Learning Units

Learning units in data science domain are heterogeneous including academic, industrial training, project experiences, and books. A common interchange format or information of how the learning units are actually constructed is not available. Therefore, an anatomy of the learning units is another key activity to design a tailored guideline. Instead of hiring a human to break them down to the learning objectives, we address this issue by facilitating crowdsourcing for the cost saving as well as for the continuous data acquisition. This project particularly focuses on the data crowdsourcing, which aims to enrich and expand the given data set.

Two HTML pages are implemented for the data acquisition from the crowd. Both pages support a rich user interface that simplifies the contribution from the user. We positioned them in the self-assessment step to force all users to make this contribution. This is also in line with the quality assurance of crowdsourcing because the more users contribute, the better service will be.

A. Level Design

The level of learning units must be determined in order to provide a step-by-step learning path. This type of task is a data labeling task. The user is requested to organize a set of learning subjects that are randomly distributed as shown in Fig 2a. Fig 2a and b are a sample HTML page that illustrates before and after user's intervention. Learning subjects are positioned in the relevant knowledge area container that is exclusive to other knowledge areas. The user is able to reposition a learning subject by a drag-and-drop action. Once the task is completed, a voting function is triggered. It increments the score for each learning subject according to its final position.

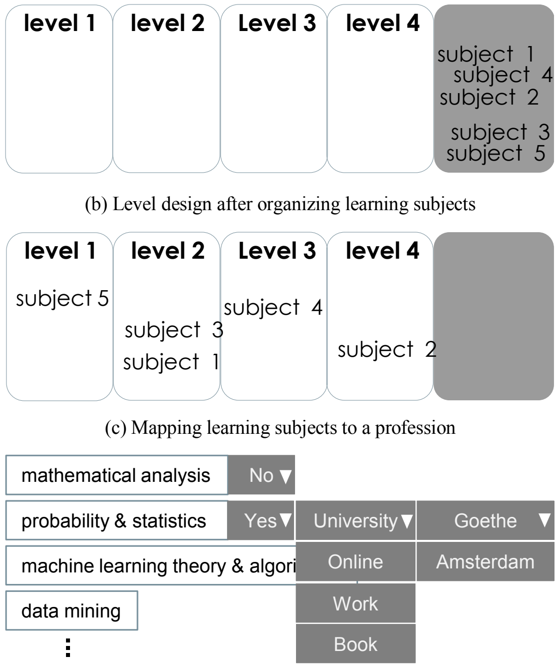
B. Mapping to Profession

Mapping a combination of the learning subjects to a profession – the current most similar profession determined in Section II.B – is a classification problem. The user is asked to select if a learning subject has already been studied. If the response is positive, one of learning methods among online, university, work or book needs to be specified as shown in Fig 2c. Either university or work is chosen, the system fetches the user data of university and company. On the other hand, one can search available online courses using EDISON inventory [2] and find a book using Google books APIs [12].

Finally, the system identifies the frequency of accomplished learning subjects per each competence level and then stores the new average results into the database. For the sake of the user-friendly service, a drop-down option field or a type-ahead prompt is provided.

Fig 2. Example user interfaces of data crowdsourcing

(a) Level design before organizing learning subjects



(b) Level design after organizing learning subjects

(c) Mapping learning subjects to a profession

More formally, k_i is a knowledge area and let l_j is a level. $\mathrm{K} = \{k_{DSDA,}\, k_{DSDM}, k_{DSENG}, k_{DSRM}, k_{DSDK}\}$ be the set of five knowledge areas and $L = \{l_{1,}\, l_2, l_3, l_4\}$ be the set of four levels. Let $f_{i,j}$ be the occurrence count of level l_j in knowledge area k_i . For instance, $f_{DSDA,2} = 3$ is interpreted as a user has completed three learning subjects of level two in the data analytics knowledge area.

In addition, let p_i be a profession in data science domain, where $P = \{p_{1,}\, p_2, \dots, p_k\}$ is the set of professions in data science domain and the current number of existing professions is nineteen. The average function $A_{i,j}$ of level l_j in knowledge area k_i for *N* total number of contributors is therefore

$$A_{i,j} = \frac{1}{N}\sum_{i=1}^{N} f_{i,j}$$

Finally, a profession p_k is described as a matrix. Each profession consists of twenty cells, a unique set of numerical value.

$$p_k = \begin{bmatrix} A_{DSDA,1} & A_{DSDA,2} & A_{DSDA,3} & A_{DSDA,4} \\ A_{DSDM,1} & A_{DSDM,2} & A_{DSDM,3} & A_{DSDM,4} \\ A_{DSEN,1} & A_{DSEN,2} & A_{DSEN,3} & A_{DSEN,4} \\ A_{DSRM,1} & A_{DSRM,2} & A_{DSRM,3} & A_{DSRM,4} \\ A_{DSDK,1} & A_{DSDK,2} & A_{DSDK,3} & A_{DSDK,4} \end{bmatrix}$$

C. *Design a Learning Path*

p_{to-be} is the desired profession consisting of twenty real numbers in a five by four matrix. Similarly, p_{as-is} is the current most similar profession. The missing gap is detected by the *Gap function* which is a key function for designing the tailored learning path.

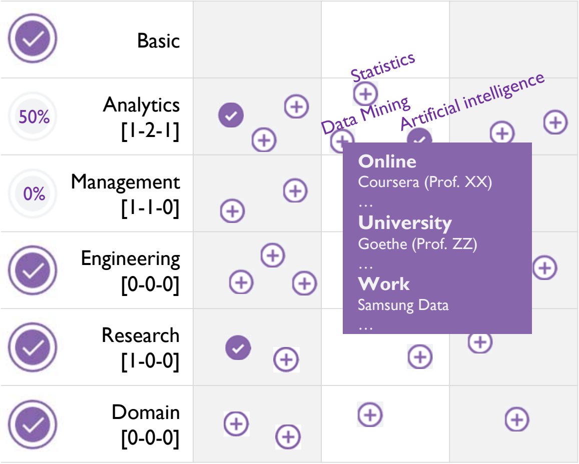
$$Gap = ROUND(p_{to-be}) - ROUND(p_{as-is})$$

The result is a five by four matrix consisting of integer numbers due to the rounding. The negative number is replaced by zero because negative means that practitioner already exceeds the competence of the dedicated profession. Finally, the non-negative integers are capable of indicating the deficient knowledge areas and how many subjects need to be completed. For instance, we can create the tailored learning paths shown in Fig 3 according to the Gap function below.

$$\begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 4 & 2 \\ 1 & 2 & 2 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 2 & 2 & 1 \\ 0 & 1 & 1 & 0 \\ 2 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

The first level of all learning subjects is a mandatory path and they are depicted at Basic area. Besides this, each knowledge area consists of three levels. Learning subjects appear at the corresponding level and knowledge area. When a user clicks a learning subject, all possible learning units are shown in a popup.

Fig 3. Example learning path. A tailored learning path is created according to the Gap function.



IV. Gamification

Beyond the traditional education, we aim to provide an entertainment-like learning experience by facilitating game design metaphors. This technique is called gamification and it is growing in popularity in the education domain. Gamification is used as an approach to improve a practitioner engagement aids achievement the learning goal through higher motivation. The used game elements in our work are Skill tree, Nonlinear learning path, Storylines, Rewards, and Character. Each element is addressed in the subsections.

A. *Skill Tree*

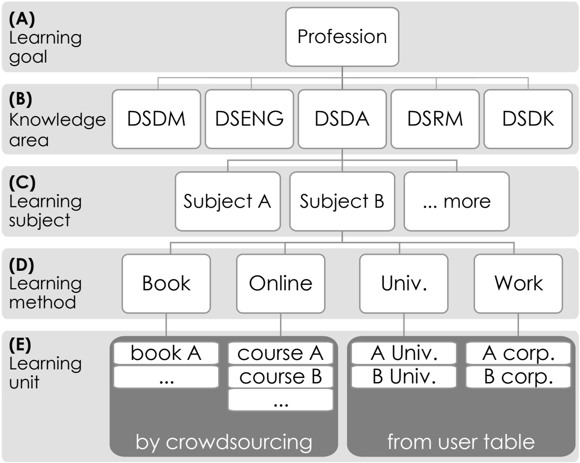
The skill tree is the way of organizing information with a hierarchy that allows one component to build on another. Components are deactivated by default, but one can unlock them after completing the required prerequisite. In other words, a practitioner should master a component that demonstrates lower-level knowledge before they move on to the more advanced one that demonstrates higher-level skill. This learning concept is following bloom's taxonomy and it is called mastery learning.

Two types of skill trees is provided based on Fig 4. The former skill tree represents a quick overview ranging depth (A) to (C). A practitioner in this view is capable of checking the status of the learning progress as well as identifing the relationship between knowledge area and their learning subject. The latter one depicts a more detailed view of data science education in five depths ranging (A) to (E). The last depth is the leaf nodes consisting of learning units that give a concrete hint what user exactly can do. It is noteworthy to mention that the view with five depths contains directional information at depth (C). The direction is key information to provide a learning path that determines what is prerequisite and what is the following learning subject.

B. *Nonlinear Learning Path*

The learning path is a network, not a line. This is the way that a human naturally learns something. We open multiple

Fig 4. Tree structure view in data science education



possibilities to achieve a learning objective through four learning methods as shown in Fig 4(D). For instance, one may take an online course to master a learning subject, whereas the other may read a relevant book. Note that a learning unit can have a redundancy. In other words, a single learning unit can appear under multiple learning subjects.

In addition, a new learning path is optionally available for the one who pursues to achieve a higher learning goal. A clickable button appears for the one who has achieved the initial learning goal. For instance, the system gives a recommendation to be a data scientist for the one who has completed the learning path to be a data analyst. This feature allows users to keep engaging in learning. In order to implement such feature, we harness cosine similarity function mentioned early in Section II.B. It measures the similarity between two professions as shown in TABLE I. For instance, the most similar profession to the data analyst is the business analyst. The result is identified empirically by assigning each data point of data analyst and the rest profession into the cosine similarity function. Moreover, the system allows one to skip a number of subjects or knowledge areas. Nevertheless, it provides an optional test session if the practitioner wants to check their correct proficiency of knowledge.

C. Storylines

It is good to have a storyline, not only because the user would not be lost in the service world, but also it helps to build a tight association with the service. In general, one can feel a sense of contribution to the story because it is developed around the activities of users. Introduction video and quests are implemented to meet this objective. Introduction video is a user-friendly guideline to introduce an objective of the service. The video aims to bring a simple and clear message to the user in a short viewing time. This helps one to define what he or she can benefit from the service. In general, the first quest is given to the user at the end of the video clip.

Quests are a popular technique for the reinforcing engagement. It provides users with a sense of accomplishment. The completion of certain quests brings a sense of achievement. Side quests are not mandatory, but optional tasks for the one who wants to add a specialty on standard competence. The examples could be an internship or study data science abroad as shown in Table III at the record number 11.

Table III. Example Quests. Quest is a popular technique for the reinforcing engagement.

no	Quest	Type	Level
1	User profile creation	Tutorial	Easy
2	Complete the questionnaire	Tutorial	Easy
3	Assess my competences	Tutorial	Easy
4	Get another assessment by clicking	Tutorial	Easy
5	Select an reward that you want to achieve	Tutorial	Easy
6	Achieve an reward that you have selected	Motivation	Ok
7	Complete three courses	Motivation	Ok
8	Complete three knowledge areas	Motivation	Ok
9	Take a course together with another user	Motivation	Easy
10	Update the questionnaire for every six month	Tutorial	Easy
11	Study data science abroad	Optional	Hard
...	...	...	...

D. Rewards and Recognition

A sense of being rewarded or being recognized is an obvious factor to reinforce engagement. These elements provide one with a sense of fame. The leaderboard is an effective way to show a user quickly where they currently stand among others. For instance, every Monday the system releases new ranking on the board and shows the comparison to one's friends. From a user standpoint, it is recommended to make this result visible on the site to motivate users to move up on the leaderboard.

E. Character

There are two characters developing the storyline. One is a player character in a blue suit, another is a non-player character in a red suit. The latter character is guiding the user to attract to the game world, also known as a service world. It provides an initial motivation to use the service, and it plays a role like a mentor to help a practitioner to meet the learning goal.

TABLE IV. Example Rewards. A sense of being rewarded or being recognized is an obvious factor to reinforce engagement.

no	Reward	Description
1	Concentration	complete an unit without surfing internet
2	Punctuality	complete an unit for a predefined time slot
3	Attendance	take a lecture everyday until accomplish
4	Best Grade	complete a course with the best grade
5	Completion	a skill branch has been completed
6	Humanism	complete a course together with another user
7	Hidden	an easter egg. It is unlocked under a certain condition
...	...	...

V. Conclusion

This paper proposes a tailored guide to construct a training program for those who want to become a data scientist or a join a related profession in data science. Nineteen professions were identified by applying EDISON framework, whereas learning units and its relation to the professions has been identified by a contribution from all participants.

A noteworthy feature is that we used gamification technique with five game metaphors. We aim to guide a complex topic to the practitioner with a strong engagement. We are currently implementing it as a web service. The current alpha version is available under goo.gl/XJAt8c (to be updated)

Acknowledgment

This paper is supported by the project EDISON within the Horizon 2020 Framework Programme of the European Union under Grant No. 675419. We would like to thank Giulia Patucca, Martino Palladini, Mattia Dallago, Carlo Riccucci and Petriccione Pierpaolo at ENGINEERING for the support and implementing of the web application. We also gratefully acknowledge Matthew Forrest, Todor Ivanov for his support for this work and kind advice. We also thank the anonymous reviewers for their helpful comments.

References

[1] M. G. Institute, S. Lund, J. Manyika, S. Nyquist, L. Mendonca, and S. Ramaswamy, *Game changers: Five opportunities for US growth and renewal*. McKinsey Global Institute, 2013.

[2] "Project deliverables | Edison Project." [Online]. Available: http://edison-project.eu/project-deliverables. [Accessed: 05-Sep-2016].

[3] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in Data Crowdsourcing," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 901–911, 2016.

[4] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller, "TurKit: Human Computation Algorithms on Mechanical Turk," in *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2010, pp. 57–66.

[5] K. M. Kapp, *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons, 2012.

[6] A. Iosup and D. Epema, "An Experience Report on Using Gamification in Technical Higher Education," in *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, New York, NY, USA, 2014, pp. 27–32.

[7] S. de Sousa Borges, V. H. S. Durelli, H. M. Reis, and S. Isotani, "A Systematic Mapping on Gamification Applied to Education," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, New York, NY, USA, 2014, pp. 216–222.

[8] "Writing Learning Objectives For eLearning: What eLearning Professionals Should Know," *eLearning Industry*, 31-Aug-2015. [Online]. Available: https://elearningindustry.com/writing-learning-objectives-for-elearning-what-elearning-professionals-should-know. [Accessed: 05-Sep-2016].

[9] "European e-Competence Framework." .

[10] R. Nikolov, "A Model for European e-Competence Framework Development in a University Environment," in *Stimulating Personal Development and Knowledge Sharing*, Bulgaria, 2008.

[11] S. Bohlinger, "Competences as the Core Element of the European Qualifications Framework," *Eur. J. Vocat. Train.*, vol. 42, no. 1, pp. 96–112, 2008.

[12] "Google Books APIs," *Google Developers*. [Online]. Available: https://developers.google.com/books/. [Accessed: 28-Sep-2016].