Homo Ludens 1(6)/2014

Gamifying science – the issue of data validation

PAWEŁ KLEKA, PAWEŁ ŁUPKOWSKI
Adam Mickiewicz University, Poznań

Abstract

In this paper we will present the idea of gamifying scientific processes. Game design techniques appear to be very useful when we are dealing with processing enormous amounts of scientific data (think of domains such as biology or astronomy, for example). We will focus our attention on the issue of the effectiveness of such an approach, as well as on the issue of procedures and techniques that ensure the high quality of results obtained via gamified scientific processes.

KEYWORDS: *gamification, scientific data processing, games with a purpose, data validation*

1. Introduction

In recent years we have been observing a growing popularity of gamification (cf. e.g. Deterding et al., 2011) in different spheres of life. We witness this trend while entering schools, universities, companies and in our everyday lives. What is interesting, gamification appears also to be a convenient solu-

tion for – broadly understood – problem solving. In this domain, examples of gamification might be gathered under one label, namely the Games With a Purpose (GWAP). The term was coined by Luis von Ahn (2006), and also the first successfully implemented GWAP was of his authorship.

One thing should be stated here – GWAPs are not serious games. Serious games are generally aimed at changing players' behaviour or attitudes (see Connolly et al., 2012), while GWAPs aim at fun and amusement. From the player's perspective it is enjoying the game that counts. The game, however, is designed in such a way that the player solves certain problems as a 'side effect' of playing. It seems important that it is not necessary for the player to be aware of the hidden function of a GWAP game (however it might improve his/her motivation). As Luis von Ahn puts it: "A GWAP ... is a game in which the players perform a useful computation as a side effect of enjoyable play" (von Ahn, Dabbish, 2008). Games like these are used, e.g., for image labelling to improve search engines and accessibility of web pages (e.g., *ESP* game, *Squiggl* – see von Ahn, 2006) or gathering common sense data about everyday objects (e.g., *Verbosity* – see von Ahn, Dabbish, 2008).

2. Gamifying science

Interestingly, GWAPs might be (and are) successfully used in the scientific domain (see Bowser, Hansen, Preece, 2013 for a brief overview) We may roughly divide GWAPs' applications in this domain into two categories:

- GWAPs that help to gather new scientific data;
- GWAPs that help to analyse already existing enormous amounts of scientific data (of various sorts).

Both categories have one common feature: they engage non-experts into a scientific process (what is often referred to as 'citizen science' – see Bowser, Hansen, Preece, 2013).

An interesting example of a (very successful) game of the first category is the *Foldit* game (<http://fold.it/portal/>). It is designed to help to solve a certain problem in the field of biology. The problem is to establish possible structures of proteins. There are an enormous number of ways in which a single protein can fold. As we may read on the website of the game: "*Foldit* attempts to predict the structure of a protein by taking advantage of humans' puzzle-solving intuitions and having people play competitively to fold the best proteins" (<http://fold.it/portal/info/science>). In *Foldit* players not only predict possible protein structures but can also design brand new proteins. The game interface is presented in Figure 1.

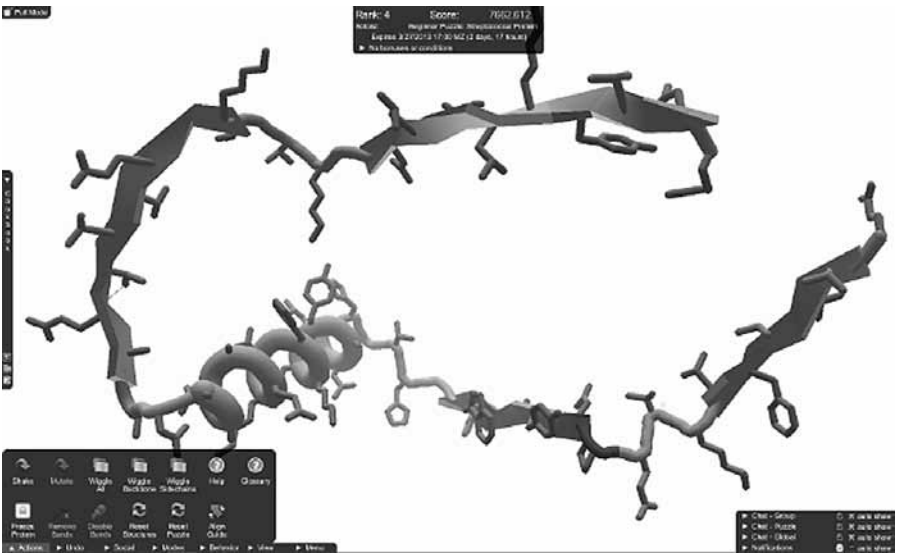

Figure 1. The *Foldit* game (source: <http://fold.it/portal/>)

Another example of such a game may be *QuestGen*. The game design was described by Paweł Łupkowski (2011) and its implementation is currently under testing. The idea of the game is to engage players in generating a large collection of questions for a certain piece of story written in natural language. The collection along with the stories will be then used as input in the research on question processing (see Wiśniewski, 2013).

In this paper we will be more interested in GWAPs of the second category. After Cooper et al. (2010) we may refer to them as 'scientific discovery games'. A game of this type is intended to help to process large amounts of scientific data of various sorts. The main tasks performed by human players in this case are intelligent data analysis and classification tasks. That is why this type of gamification in science is attractive. Clear examples of such games are *GalaxyZoo* (Masters et al., 2010) and *Wardorbe* (Venhuizen, Basile, Evong, Bos, 2013).

In *GalaxyZoo* <http://www.galaxyzoo.org/> users classify pictures of galaxies obtained from *Sloan Digital Sky Survey* (SDSS). An additional motivation for players is that most of the pictures have not been seen by anybody before them (as the slogan of the game says: "Few have witnessed what you're about to see"). Each galaxy is classified as belonging to one of the categories that are clearly recognizable in the game interface (see Figure 2). Importantly, the game is very intuitive and only a short introduction is needed to start playing. The first edition of *Galaxy Zoo* was so successful that now we can play the second edition of this game (with more advanced classifications available).

Galaxy Zoo is now also a part of a broader project called *Zooniverse* <https://www.zooniverse.org/> which gathers in one place games of similar character (for example, we can classify pictures of the Moon surface in *MoonZoo*, classify objects on pictures of the sea floor in *SeaFloor Explorer*, or even annotate and tag diaries from the First World War in *Operation War Diary*).

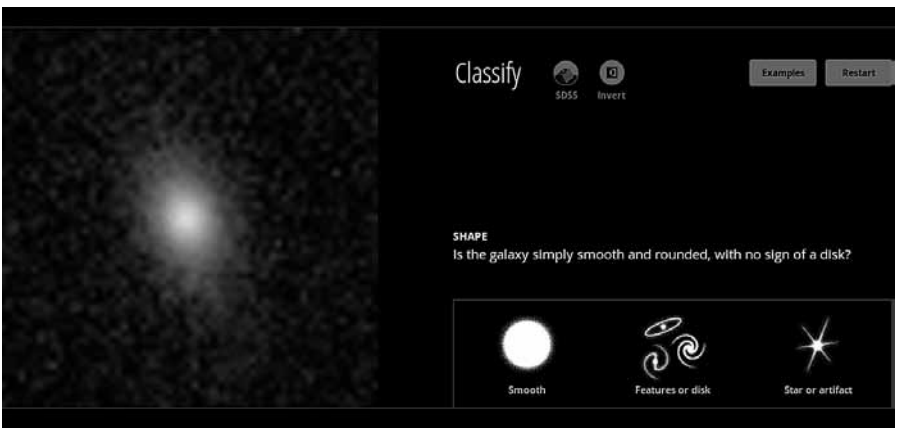

Figure 2. *Galaxy Zoo* interface (source: <http://www.galaxyzoo.org/>)

Wordrobe <http://wordrobe.org> is a set of games developed to enable semantic annotation of the natural language data from the *Groningen Meaning Bank* (GMB) <http://gmb.let.rug.nl/>. Each game aims at a different level of annotation. For instance in *Senses* the player's goal is to identify the correct sense of a word, whereas in *Others* the player has to identify to whom (or what) the word "others" refers in a piece of text. An exemplary task from *Others* game is presented in Figure 3. All tasks in the games are obtained from GMB and the annotated texts are used to improve the corpus.

It is worth considering the motivations for using a gamified approach to scientific data processing. When we are facing the problem of analysing enormous amounts of scientific data (like natural language corpora or pictures database as discussed above), we may think of outsourcing this problem. One of the most natural ways to do this is to employ some experts to do the task. However, this solution involves time and costs (usually high). If we want to speed up the process and lower the number of involved specialists we may use an outsourcing platform, at which we delegate small parts of the task to other people (non-experts). For this purpose one may use dedicated platforms like *Amazon Mechanical Turk* (<https://www.mturk.com>) or *CrowdFlower* (<http://crowdflower.com/>). The gamified approach allows for further savings in terms of costs because we do not have to pay our non-expert players

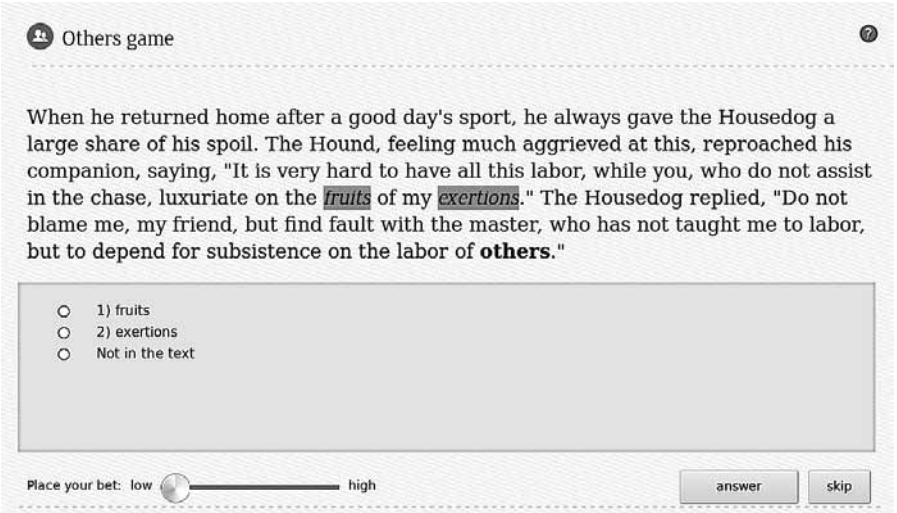

Figure 3. An exemplary task from the game *Others* (source: <http://wordrobe.org>)

in a GWAP setting. What is more, a problem presented in a game-like form might be more attractive to non-experts and may gather more of them.

One may also ask: why bother with outsourcing the problem to humans? Maybe it would be simpler to use some computer methods. With the increasing speed and computing power, computers replace humans in many tasks. Fewer and fewer tasks related to data analysis are performed by people, but still people are better than computers in lexical tasks involving the recognition of emotions, decoding synonyms, reasoning context, organizing events and assigning meanings (Venhuizen, Basile, Evong, Bos, 2013). People are also superior to automatic programs in graphics tasks, such as recognizing similarities, extracting figures from the background and decoding meanings (Darg et al., 2010).

3. Effectiveness and validity of the results obtained with GWAPs

What the games presented in the previous section have in common is that there are both scientific data and non-experts (in a role traditionally ascribed to the group of experts). What is more, data processing is done in a game (or game-like) environment. The two following questions seem to be quite natural. Firstly, how effective is gamified analysis of scientific data? Secondly – and even more importantly – how valuable is the output data of such gamified processes?

Let us start with the effectiveness of GWAPs. Thanks to people's motivations associated with playing a game, the application of GWAPs can bring a stunning number of participants and their voluntary engagement. The first edition of *Galaxy Zoo* gathered more than 100 000 people who classified more than 300 000 galaxies, with an average of about 30 classifications per player (Lintott et al., 2008, p. 1188). The numbers for other GWAPs gathered within the *Zoouniverse* project are also impressive[1]: *Planet Four* (Mars images) gathered 115 703 participants with 4 335 094 classifications made; *Moon Zoo* has 3 773 752 classifications; and *SeaFloor Explorer* has 2 177 030 classified images of sea floor. In the case of *Wordrobe* – described in the previous section – the numbers are smaller but the project is young and tasks are more demanding. 962 players and 41 541 annotations were noted after a few months since the launch of the game (as reported in Venhuizen, Basile, Evong, Bos, 2013).

As might be noticed, the potential is enormous. There is only one question left – how to ensure that the data we obtain via gamified systems is valuable?

To ensure reliability of such amounts of data produced as a side effect of playing GWAP games, designers employ many techniques. First of all, all the games described demand each user to register before he/she can contribute. Players are also informed what the game is designed for (the intention is to decrease the number of players who want to "game the system" – see Bowser, Hansen, Preece, 2013). Many in-game mechanisms are also employed, like users testing. For example, in *Galaxy Zoo* the first task of the user is to classify pictures of already known galaxies (with established classifications). Another option is task repetition (the same task is repeatedly presented to different players and the consistency of solutions is checked). Results obtained via GWAPs are also compared and validated with results of automated techniques and those provided by experts. And thus *Foldid* output data are compared with data obtained by *Roseta's* rebuild and refine protocol and analysed by experts (Cooper et al., p. 757). Also, *Galaxy Zoo* output data are confronted with the data from automatic classifiers and expert knowledge (Lintott et al., 2008, p. 1183).

When it comes to scientific discovery games and the valuation of the data obtained, we may also make use of some simple statistical methods. Being aware of these methods is important for designing a GWAP in scientific domain. In the case of classification tasks that are exploited by the games described (like *Galaxy Zoo* or *Others*), examining only the inter-rater agreement within a non-experts group will not answer the question of validity.

[1] All the data from the web pages of the project have been retrieved on 21 July 2014.

The reason for this is the number of the subjects – the coefficients of the inter-rater agreement depend on the sample size, which in GWAPs can be very large.

The first step of solving such a problem is using quantitative notations. If players can make notations only on a dichotomous (yes/no) scale, we cannot use parametric statistics. Only implementation of ordinal (measuring in terms of "less/equal/more") or numerical scales (with units, e.g., per cents) enables us to apply parametric statistics in analysing the quality of the obtained results. For example, regression analysis with residual analysis could be used to discover the players with results that protrude from the average scores because of giving random answers or trying to "game the system" (instead of following the game's main objective). An example of such a quantitative notation in use might be the 'bet' element in the *Others* game (see Figure 4). After choosing your answer you bet on how sure of the answer you are.

The second step is to estimate the equivalence of answers given by experts and non-experts. Taking the inter-rater agreement coefficient as a starting point, we can determine how many non-experts have given answers with the same quality as one of the experts. One example of such an analysis is related to a project of human linguistic annotation (Snow, O'Connor, Jurafsky, Ng, 2008). For the textual material in which players had to recognize one of the six basic emotions, two non-experts performed better than, or just as well as, one expert at recognizing *sadness*, *anger* and *disgust*. For *joy* 7 non-experts were needed, whereas the number for *surprise* was 8, and for *fear* more than 10 non-experts were required to provide results equal to one expert. This example shows that it is possible to estimate how many non-experts are needed to provide solutions as good as those provided by experts for a certain task.

In the GWAP context we could hardly force experts to evaluate all the data. In order to avoid that we can use methods based on the so-called gold standard test. In this test a group of experts classify parts of the data, and the results are used to assess the quality of data obtained from non-experts. Usage of this method has been reported for many GWAP systems (e.g., Lintott et al., 2008, p. 1183; Venhuizen, Basile, Evong, Bos, section 4). The kind of estimate described in the previous paragraph is useful for establishing the gold standard test in classification tasks.

In the third step we can use a simple procedure of weighing non-experts' responses in the relation to the gold standard to analyse the answers provided by the players. In order to do this we may apply the following formula:

$$w = \log\left(\frac{V}{1-V}\right)$$

where: w – weight, V – validity:

$$w = \log\left(\frac{ci}{ci + er}\right)$$

where: ci – correct, er – incorrect classification.

If the degree of agreement between non-experts and experts is close to 50%, it will be treated as guessing and thus it will present no value [log(0.5) = 0]. If the agreement is less than 50%, we can assign a negative weight, and if the accuracy is higher than the random 50%, a positive weight can be assigned. In this way we are able to distinguish good, worse and very bad solutions provided by non-experts.

In the fourth step we may consider using a more sophisticated matrix to classify the players' answers. Such a full response matrix contains the incorrect rejection (marking good results as bad) and correct rejection (marking bad results as bad), as seen in Table 1.

Table 1. The matrix of all possible answers

	Not pass	Pass
Correct	False positive	True positive
Incorrect	True negative	False negative

Source: Tanner Jr., Swets, 1954

Let us define 'sensitivity' (recall ratio) as the ratio "True positive / (True positive + False negative)" and 'precision' as the ratio "True negative / (True negative + False positive)". With these terms we may use Receiver Operating Curves to visualize the quality of the classification in the data set. The theoretical plot is shown in Figure 4.

The closer the result is to the upper left corner, the better prediction it implies. The worst of the three results belongs to the player X_2 and lies on the random guess line (the diagonal line). Her/his accuracy is 50%. The player X_1 has better accuracy than the player X_3, which additionally assigns categories in the opposite way. But still X_3 gives better predictions than X_2.

4. Summary

In modern science there are frequent problems with processing enormous amounts of data. As we intended to show in this paper, a gamified approach (via GWAPs) to such a problem brings a very effective solution. What is more,

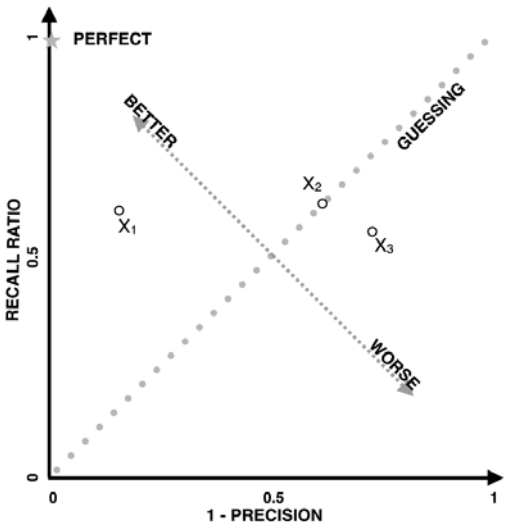

Figure 4. The Receiver Operating Curves space plot

games with a purpose allow for obtaining valuable data, while employing non-expert players to perform a part of scientific process in a game-like environment. Such a validity might be achieved by the use of design techniques and universal statistical techniques as it was discussed in our paper.

REFERENCES

Ahn von, L. (2006). Games with a Purpose. *Computer, 39*(6), 92–94.

Ahn von, L., Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*(8), 58–67.

Bowser, A., Hansen, D., Preece, J. (2013). Gamifying Citizen Science: Lessons and Future Directions, In: *Proceedings of the International Conference on Human Factors in Computing Systems*. CHI 2013, Paris, France.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., Boyle, J. M. (September 2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., Foldit Players (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

Darg, D. W., Kaviraj, S., Lintott, C. J., Schawinski, K., Sarzi, M., Bamford, S., Vondenberg, J. (2010). Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies. *Monthly Notices of the Royal Astronomical Society, 401*, 1043–1056.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., Dixon, D. (2011). Gamification: Using Game Design Elements in Non-Gaming Contexts. In: *Proceedings of the International Conference on Human Factors in Computing Systems*. CHI 2011, Vancouver, BC, Canada.

Lintott, Ch. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, Ph., von den Berg, J. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society, 389*, 1179–1189.

Łupkowski, P. (2011). Human computation – how people solve difficult AI problems (having fun doing it). *Homo Ludens, 1*(3), 81–94.

Masters, K. L., Nichol, R. C., Hoyle, B., Lintott, Ch., Bamford, S., Edmondson, E. M., Fortson, L., Keel, W. C., Schawinski, K., Smith, A., Thomas, D. (2010). Galaxy Zoo: Bars in Disk Galaxies. *Monthly Notices of the Royal Astronomical Society.* Online: <http://arxiv.org/abs/1003.0449>. Access date: 21 July 2014.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y. (2008). Cheap and fast – but is it good?: evaluating non-expert annotations for natural language tasks. In: *Proceedings of the conference on empirical methods in natural language processing* (p. 254–263).

Tanner Jr., W. P., Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401–409.

Venhuizen, N. J., Basile, V., Evong, K., Bos, J. (2013). Gamification for word sense labeling. In: *Proceeding of the 10th International Conference on Computational Semantics* (IWCS-2013; p. 397–403).

dr Paweł Kleka, Institute of Psychology, Adam Mickiewicz University, Poznań (Instytut Psychologii, Uniwersytet im. Adama Mickiewicza), Pawel.Kleka@amu.edu.pl

dr Paweł Łupkowski, Department of Logic and Cognitive Science, Institute of Psychology, Adam Mickiewicz University, Poznań (Zakład Logiki i Kognitywistyki, Instytut Psychologii, Uniwersytet im. Adama Mickiewicza), Pawel.Lupkowski@amu.edu.pl

Grywalizacja badań naukowych – jak zapewnić wysoką jakość otrzymywanych wyników?

Abstrakt

W artykule przedstawimy zagadnienie grywalizacji badań naukowych. Metody zaczerpnięte z dziedziny projektowania gier okazują się przydat-

ne w obliczu problemu przetwarzania ogromnych ilości danych otrzymywanych we współczesnych badaniach naukowych (np. z dziedziny biologii czy astronomii). Skoncentrujemy się na problemie efektywności takiego rozwiązania oraz na pytaniu o procedury i techniki, które zapewniają wysoką jakość rezultatów otrzymywanych w zgrywalizowanych badaniach naukowych.

SŁOWA KLUCZOWE: *grywalizacja, przetwarzanie danych naukowych, gry skierowane na cel (games with a purpose), wiarygodność danych*