Big Data techniques in science education and what story Google Trends tells us about Science?

M. Şahin Bülbül
Renato P. dos Santos
Isadora Luiz Lemes

ABSTRACT

The intention of this work is to provide a quick overview of what Big Data is and present a few examples of techniques through which it can contribute to Science Education. Google offers the Google Trends (GT) free analysis tool that allows users to sort through several years of Google search queries from around the world to get a graphical plotting showing the popularity of chosen search terms both over region and time. According to the time, region, and frequency of search, three kinds of data are evaluated in terms of compatibility with a sort of "correlation analysis". A few techniques of extracting meaning from them are exemplified through geographical searches for 'Solar Eclipse' in USA and through temporal searches of the term 'research' in the period 2013-2017. In addition, and as the main study, an experiment was conducted to replicate with Big Data and GT Taşdere, Özsevgeç, and Turkmen's survey on the Nature of Science (NoS). To that end, the same nine concepts they selected were searched in GT. Two-way correlation analysis was performed on these words, and those pairs with a Pearson Correlation of 0.8 and higher were used to build a conceptual network. Three main levels emerge in our hierarchical conceptual network and, as a result of this structuring, a *storytelling* can be built: Science is seen, in a more publicly understandable level, as associated to 'laws', followed by a less-visible level of research being associated to 'building theories', and then, in a even lesser understanding level, the scientists doing experiments to test hypotheses, which are confirmed or not by observation – an image of scientists' work shaped in a large degree by popular media.

Keywords: Science Education, Big Data in Education, Google Trends, learning-with-Big-Data, educational technologies.

M. Şahin Bülbül holds a PhD degree (Physics). Presently, he is Associate Professor at Kafkas University and Researcher in Physics Education, Science and Special Education, Chaos and Education, Big Data, and Creativity. Address: Kafkas Üniversitesi, Faculty of Education, Kars, Turkey. E-mail: msahinbulbul@gmail.com
Renato P. dos Santos holds a ScD degree (Physics). Presently, he is Associate Professor at ULBRA – Lutheran University of Brazil, Doctoral Program in Science and Mathematics Education. Address: ULBRA/PPGECIM, Av. Farroupilha, 8001, pr. 14, s. 338, 92450-900 – Canoas/RS. E-mail: renatopsantos@ulbra.edu.br
Isadora Luiz Lemes holds a BSc degree (Physics). Presently, she is a graduate student at ULBRA – Lutheran University of Brazil, Doctoral Program in Science and Mathematics Education. Address: ULBRA/PPGECIM, Av. Farroupilha, 8001, pr. 14, s. 338, 92450-900 – Canoas/RS. E-mail: isa.ulbra@hotmail.com
Received for publication on 8 Nov. 2017. Acepted, after revision, on 30 Nov. 2017.

Técnicas de Big Data na educação científica e qual estória[1] o Google Trends nos conta sobre Ciência?

RESUMO

A intenção deste trabalho é fornecer uma rápida visão geral do que é Big Data e apresentar alguns exemplos de técnicas através das quais Big Data pode contribuir para a Educação em Ciências. O Google oferece a ferramenta de análise gratuita do Google Trends (GT) que permite aos usuários classificar vários anos de consultas de pesquisa do Google de todo o mundo para obter um gráfico que mostre a popularidade dos termos de pesquisa escolhidos tanto na região quanto no tempo. De acordo com o tempo, região e frequência de pesquisa, três tipos de dados são avaliados em termos de compatibilidade com uma espécie de "análise de correlação". Algumas técnicas para extrair o significado delas são exemplificadas através de pesquisas geográficas para "Eclipse Solar" nos EUA e através de pesquisas temporais do termo "*research*" (pesquisa) no período 2013-2017. Além disso, e como estudo principal, um experimento foi conduzido para replicar com Big Data e GT a pesquisa de Taşdere, Özsevgeç e Turkmen sobre a Natureza da Ciência (NoS). Para esse fim, os mesmos nove conceitos que eles selecionaram foram pesquisados em GT. A análise de correlação bidirecional foi realizada para essas palavras e aqueles pares, com Correlação de Pearson de 0,8 e superior, foram utilizados para construir uma rede conceitual. Três níveis principais emergem em nossa rede conceitual hierárquica e, como resultado dessa estruturação, uma narrativa pode ser construída: a ciência é vista, em um nível mais compreensível ao público, associada a "leis", seguida de um nível menos visível de pesquisa associada a "construção de teorias", e, em um nível de compreensão ainda menor, os cientistas realizam experimentos para testar hipóteses, confirmadas ou não pela observação – uma imagem do trabalho dos cientistas moldada, em grande medida, pelos meios de comunicação.

Palavras-chave: Ensino de Ciências, Big Data em Educação, Google Trends, aprender-com-Big-Data, tecnologias educacionais.

INTRODUCTION

Today, computers, tablets, and mobile phones are sources of digital data, and the quantity produced and processed by these sources is increasing day by day. When one thinks about the number of likes made in a second in social media, the contents of uploaded photographs and videos, it is evident that the type of analysis to be done on this massive amount of data has to be different from the ones that were used before. The science branch that has been recently seen accounting for these increasing quantities and varieties of data is the Big Data Science. For example, Big Data studies have been conducted in Healthcare by insurance companies have being doing research to optimise premiums based on client's physical activity as measured by their mobile phone (MURGIA, 2017).

While traditional data analysis is mostly done on structured data, the diversity of data in Big Data analysis is also increasing, so that the size (capacity) of the storage location of the data and the processing speed of the application to be processed have also increased. Insufficient storage space leads to incomplete data analysis, and low

[1] Sentimo-nos obrigados a utilizar aqui a palavra 'estória', por paralelismo com o termo *storytelling* (KUMAR et al., 2008), consagrado em Ciência de Dados.

processing speed causes data accumulation. Moreover, if there are restrictions on the diversity of the data, the meaning may be lost while mining the data.

The primary concerns of Big Data research are about storing voluminous data from Big Data sources and making them meaningful by appropriate analyses. Various companies provide free software and/or data that can be stored and analysed. For example, Google Inc. makes available data from billions of monthly searches through the free app *Google Trends* (GT)[2].

GT shows the search rates of terms on the World, and trends in these rates can be followed. This app also allows one to download the search frequencies data in the desired region and time for each searched word for later analysis. Such follow-ups give preventive information. For instance, the number of cars that are increasing rapidly on a given road tells us that there seems to be a traffic jam on that same road in a very short time ahead. Likewise, if the number of searches for a particular term is rapidly increasing, one can assume there must be some critical event related to it.

GT provides quantitative values that have to be interpreted on the basis of the concept of "search," but social media data can provide more reliable and in-depth interpretation (PROVOST; FAWCETT, 2013). For example, pictures and videos that are shared through social media can be understood through different techniques. The intelligent software may interpret the contents of pictures and video, analyse who likes and responds to them, build a better understanding of these users, and be able to predict and trigger possible events accordingly. Therefore, a user who likes and shares every single image of a specific soccer team is probably responsive to anything about this team, and the software might suggest him/her a group in which enthusiastic fans come together.

Thousands of scientific works have been made using GT in various areas of knowledge. For recent representative examples, consider Public Health (CAVAZOS-REHG et al., 2015; NSOESIE; BROWNSTEIN, 2015), Economics (HEIBERGER, 2015; SCOTT; VARIAN, 2014), Education (YIN et al., 2013; ZHANG et al., 2015), and Politics (BANTIMAROUDIS, 2015; SINCLAIR; WRAY, 2015), among many others.

More in line with the scope of the present study, Guo, Zhang, & Zhai (2010) have used GT for the study of human curiosity, understood as a desire to acquire new information and knowledge, and its measurement, while Segev & Baram-Tsabari (2009a, b) and dos Santos (2016) have used GT to explore the public interest in Science. Furthermore, Pence and Williams (2016) recommend that all data sources were examined by Big Data studies and that Chemical Education studies should be based on already collected data instead of collecting new data.

All these problems considered, companies are making serious investments in Big Data Science. It is too early to say that the same investments are made in the field of

[2] http://www.google.com/trends/

Education, however. Only epidemic diseases are being pre-determined, and precautions are being taken, even if the course attendance of the students across the country should also be analysed.

As we believe that Big Data studies are critical in the field of education, especially Science Education, we will detail here three examples of application using GT, ranging from a mere comparison of search rates of a few terms, passing through a study of the evolution of the searches for a concept along the history, to a more in-depth correlation analysis of terms leading to the construction of a hierarchical network of concepts, from which a *storytelling* emerges.

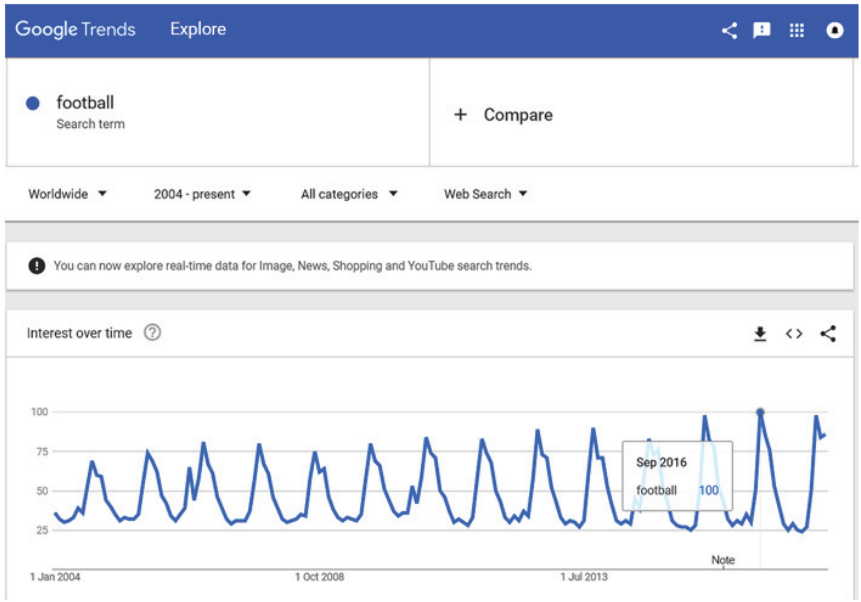
The applications of GT in Science Education presented here could be understood within the *enquiry-based learning* framework (DOSTÁL, 2015), as they involve a form of active learning based on the students posing their own questions and problems of interest and interpreting the results and answers provided by GT. We understand *learning-with-big-data* as a promising new way to learn Science (DOS SANTOS, 2015) that intends to recover the "intellectually adventurous side" of discovery learning usually lost within the so common "excruciatingly boring" (PAPERT, 2000) experience of being presented to established skills and facts (e.g., denatured laboratory practices and 'physical laws').

GOOGLE TRENDS AS A BIG DATA PROCESSOR

Google Search engine stores about one hundred billion Web searches monthly, all identified by time and place of origin. This data is continuously analysed to be used by its highly profitable advertising programs, such as *Google AdWords, DoubleClick*, *Google Analytics*, and *Google AdSense*, from which comes 90% of Alphabet Inc. (the now parent company of Google Inc.) revenue (ALPHABET INC., 2017). Fortunately, this stored information has also been made available by means of various public analytical tools released the last few years, such as *GT* and *Google Correlate*.

GT allows users to sort through several years of Google search queries, since 2004, from around the world or a particular country or State where the data is collected. The interested user has only to type the desired word e.g. 'football' in the search box and a graphical plotting (Figure 1) is drawn showing the evolution of popularity of the searches by the term over time (GOOGLE INC., 2012b). This example shows that the searches for 'football' have some periodicity, with maxima usually occurring in September each year.

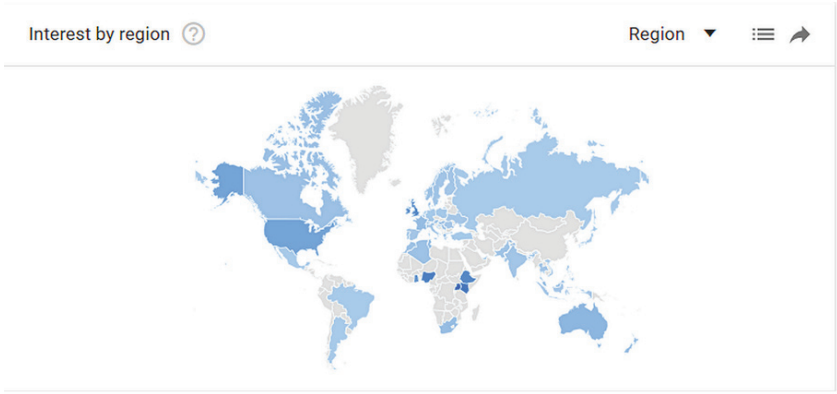
FIGURE 1 – Evolution along time of the searches in Google by the word 'football'.



Source: Google Trends (www.google.com/trends).

GT also generates maps (Figure 2), in which those regions of the world in which the searches by the word 'football' were most popular during this period are coloured darker while those with lower rates are coloured lighter. Regions with low search volumes are left colourless (grey). In this example, Uganda was the country in which the term 'football' was most searched (GOOGLE INC., 2012b).

FIGURE 2 – Distribution around the world of the searches in Google by the word 'football'.



Source: Google Trends (www.google.com/trends).

GT also provides a list of 'related topics'(Figure 3), that is, popular search terms that are similar to the one entered ('football', in this case), as well as a list of 'related topics' and a list of 'related queries', i.e., terms that are most frequently searched together with the one introduced in the same search session, within the chosen category, country, or region (GOOGLE INC., 2012b).

FIGURE 3 – List of topics related to 'football', according to searches on Google.

Source: Google Trends (www.google.com/trends).

In such analyses, it is important to take into account the language in which the search term is written, as most searches are done in local languages. For this reason, an English word with international appeal was chosen for this example; otherwise, a worldwide map could not have been generated. For example, *araştırma* means 'research' in Turkish, but if one uses *araştırma* as a search term, one will virtually get results from Turkey only and much less from Arabic countries or the rest of the world.

Another point to take into account in these analyses is that the value that appears in the ordinate of the time-charts and map graphs is not the absolute number of searches for that topic. Rather, they are normalised: each data point is divided by the total of searches done on Google Search Engine (GOOGLE INC., 2012c) on the geography and time range it represents; the resulting numbers are then scaled to a range of 0 to 100. As such, a value close to 100 indicates that it is close to the maximum value of the current search numbers for that place and time period (GOOGLE INC., 2012a). Therefore, if, e.g. at most 10% of searches for the given region and time period were for "pizza," GT would score this '100,' and all the other points, both to all searches and to other items plotted, are valued relative to it (GOOGLE INC., 2012a).

Consequently, another important issue that affects the accuracy of the analysis are the users' habits and his/her availability of access to the Internet. Individuals who are deprived of Internet access or do not have the habit of using it as much as possible prevent the results from revealing the truth. This hampering does not preclude the researcher from seeing the big picture, but it is a weakness in the method.

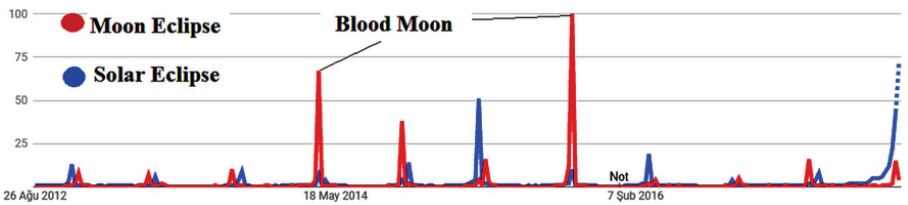
There are, therefore, three categories of essential data for science education through GT. These are positional and temporal data regarding where and when the search was done, as well as the frequency data of how often the term was searched. These 'how much,' 'where,' and 'how long' the sought-after concept or person is searched are essential for the method on which we will work.

It is also possible to select more than one word for analysis and compare their searches across region and time. Correlation between searches can be used as evidence for establishing a relationship network among concepts that have logical closeness although it does not assure a causal relationship. The underlying assumption on which this approach is based is that if both terms were frequently searched on the same occasion, there is probably some dependence among them.

FIRST EXPERIMENT: THE FULL SOLAR ECLIPSE OF AUGUST 21, 2017

Let us first compare the searches for Sun and Moon eclipses in GT. In the English language, a Sun eclipse is usually searched as "Solar eclipse," and a Moon eclipse as "Lunar eclipse." GT draws graphs in different colours to ease comparison. If one examines the frequency of these terms together, one sees that both search rates for "Lunar eclipse," displayed in red, and for "Solar eclipse," in blue, are highly correlated (Figure 4).

FIGURE 4 – Comparison between the searches for 'Solar Eclipse' and 'Lunar Eclipse' in Google.



Source: Google Trends (www.google.com/trends).

On 21 August 2017, when the full solar eclipse was to occur, the amount of searching seems to rise steeply towards the end of the graph (Figure 4). The 'bloody' (total) lunar eclipses observed in May 2014 and September 2015 increased the number of searches for these topics, as compared to other days. This situation should be considered when

there are close searches. The chosen search term, i.e., how the event and/or situation is written, is the most critical factor affecting the search results. In this example, it is necessary to consider the fact that users who do not know whether it is the Moon or the Sun eclipse may do the wrong search, and this error is reflected in the mix in the search results in Figure 4.

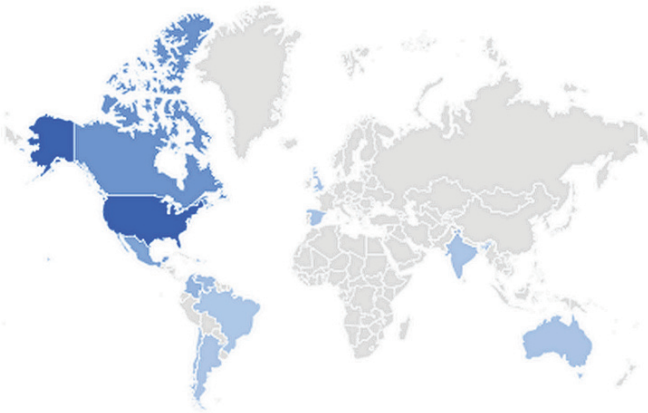
It is interesting to search for all these four spellings 'Sun eclipse,' 'Moon eclipse,' 'Lunar eclipse,' and 'Solar eclipse' simultaneously (Figure 5). Besides the language issue, which makes "Solar eclipse" more often searched than 'Sun eclipse' in English-speaking countries, there is apparently also some cultural component that makes eclipses of the Moon more searched than those of the Sun in some countries in South America, Spain, Italy, Japan, Thailand, Saudi Arabia, Kenya, Egypt, Turkey, Poland, Romania, Serbia, Denmark, and Norway.

FİGURE 5 – Distribution around the world of the searches in Google by the four spellings 'Sun eclipse,' 'Moon eclipse,' 'Lunar eclipse,' and 'Solar eclipse'.

Source: Google Trends (www.google.com/trends).

The USA is at the forefront as the country in which the full solar eclipse of August 21, 2017 will be best seen. It is, therefore, understandable that there are consequentially more searches in the USA than in other countries and this situation is reflected in the search results (Figure 6). As the number of search results increases, the related position on the World Map becomes darker than the others and the USA is shown in the darkest colour in Figure 6, while other areas appear lighter-coloured or colourless.

FİGURE 6 – Search frequency of the word “Solar Eclipse” across the world.



Source: Google Trends (www.google.com/trends).

This map can be misleading, however, because it is centred on the European continent. Researchers might ask why the search rate is also relatively high in Western Europe, Australia, and India. This issue stems from GT not providing absolute search volume raw data, but only normalised data to the total number of searches done on Google Search Engine (GOOGLE INC., 2012c) on the geography and time range it represents, as said before. This leads one to suppose that in a map centred on the US these search results would be more meaningful.

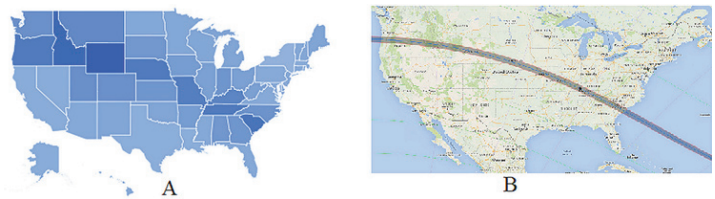
Here another factor emerges: the type of the visualisation of the data affects its interpretation. The latest stage of the Big Data collection is the display of the data, but the visualisation should be worked on before the alternatives are understood.

GT, which handles all the searches made and allows us to compare different regions of the globe, also provides a focus feature on the World Map. If one clicks on the country one wants to have more in-depth knowledge, one can see the search results restricted on the territories of the selected country. This possibility provides a more in-depth research opportunity. For example, a region of the US might be searching too often and separating the country from other nations, and one might be induced to think that there are many searches in the US as a whole. Deepening the research give ones the knowledge that the search is intensive in some but not all of the regions of that country.

When an in-depth study in the US with the word “Solar Eclipse” is made, the search frequency in all regions can be seen together (Figure 7A). The dark areas on the map are noteworthy: they resemble a strip extending from the Northwest region of the United States to the Southeast. The situation in which the search frequency builds a pattern is a condition that should be supported by other sources of information. When the regions where this solar eclipse can be best watched are examined, it is seen that they appear on the map in the same alignment (Figure 7B) as the darkened areas in the GT results. If

one had not previously known which regions where the eclipse would best be observed, he might have discovered it from the graph of Figure 7A. This previously unexpected relation may be interpreted as a *crowdledge*, that is the knowledge that emerged from Big Data analysis of individuals' digital footprints spontaneously left in the digital universe we live in by means of web searches (in this case) (DOS SANTOS, 2015).

FİGURE 7 – The frequency of the search for the word "Solar Eclipse" in the United States (A) and the regions where the solar eclipse is to be observed (B).



Source: Google Trends (www.google.com/trends).

Now, one may wonder why is the number of searches increasing when a solar eclipse is approaching. According to Scheitle (2011), the arguable assumption is that if people are interested in a particular issue, they will likely Google the web for resources, news, websites, discussion boards, and other types of information related to it. The concept of emerging patterns belongs to psychology and sociology, but their meanings constitute the field of research of "Big Data Science." This science suggests meanings by supporting the emerging patterns with different sources. While people who work with Big Data usually understand the patterns they discover, they need to confirm these patterns from the relevant field and other sources of information. This demonstrates the necessity of workers in Big Data to be field experts besides having analytical (statistical) skills, as well as computer/program use skills. GT partially covers the shortcomings of these two skills, but yet misunderstanding data visualisation can lead to incorrect results if one is not expert in the field.

Second experiment: Evolution of the searches for a concept along the history

Sometimes, one may want to follow the evolution of the searches for a concept along the history. In this case, it is useful to include the desired year to the search term under study to obtain the most relevant related concepts searched. For example, let us select the search term 'research', and select a year from 2013-2017 to do a research to reveal the queries associated with the term 'research' for each year. If this research is done for Turkey, however, uninteresting queries such as "research officers and salaries" are the more often recovered. For this reason, we decided to resume the investigation through the USA. Five searches were done, each including one of these years in front of the word "research" as in "2017 research", "2016 research", and so on. It allows one to

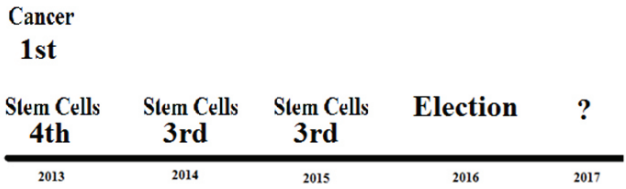
see the focus of the research done in the USA in each of the last five years. In Figure 8, one sees the predominant queries for 2013.

FİGURE 8 – Predominant queries related to the term 'research' in 2013.

Source: Google Trends (www.google.com/trends).

After the five searches for 2013 to 2017, Figure 9 shows the research topics in the USA most related to 'research'.

FİGURE 9 – Chart showing which searches according to years shows the front plan in the USA



Source: Google Trends (www.google.com/trends).

One observes in Figure 9 that 'cancer research' and especially 'stem cell research' were at the forefront from 2013 to 2015 but disappear from the top searches afterwards. On 2016, the top related query was "election", a result that, at first, might lead one to interpret that the presidential elections in the United States were influencing scientists, projects, and scientific research negatively. On second thoughts, however, this may stem from confusion between two meanings of 'research', regarding 'scientific research' and 'opinion pool', the latter quite frequent before elections. This is an example of the "secondary meanings of words" that will be discussed below. This and similar studies can be done with current scientific history studies.

In this part of our research, we used GT to reveal original examples of inquiries that can be done about science education. In the next section, we will use the technique

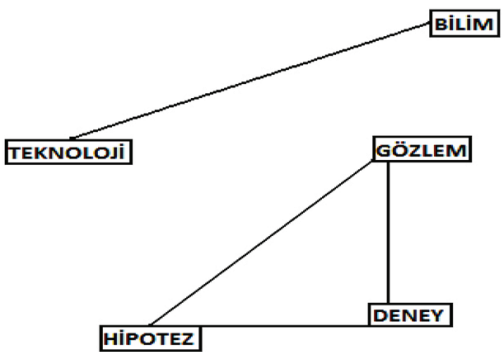
of "correlation" of GT results. It is worth ***emphasising***, however, that Big Data and GT studies provide correlation relations only, not cause-effect relationships.

Third experiment

This research aims to replicate using Big Data Science the research done with small data by Taşdere, Özsevgeç, and Turkmen (2014).

Those authors carried out a study to determine the understanding and cognitive structures for the Nature of Science (NoS) of 23 candidate teachers on the last semester of the 2011-2012 academic year from Faculty of Education Science and Technology Teacher Training Department of Uşak University. They selected the nine keywords 'science,' 'scientist,' 'experiment,' 'observation,' 'research,' 'technology,' 'hypothesis,' 'theory,' and 'law' and then asked the candidate teachers to write down the first word that came into their minds when presented to each of these, using the complementary assessment technique of Word Association Test. From the keywords and answer words collected in the pre and post-tests, the researchers drew conceptual networks according to the frequency of association intended to reveal the cognitive structures of teacher candidates. The answer words collected increased from 1170 to 1368 in the post-test, and the researchers observed that the post-test conceptual network also had a more complex and interrelated structure than the pre-test one, after the natural sciences and science history course. The analysis of the networks revealed separate relations between the concepts of science (*bilim*) and technology (*teknoloji*) and between observation (*gözlem*), hypothesis (*hipotez*), and experiment (*deney*) (Figure 10).

FİGURE 10 – Final test concept map of science teacher candidates.



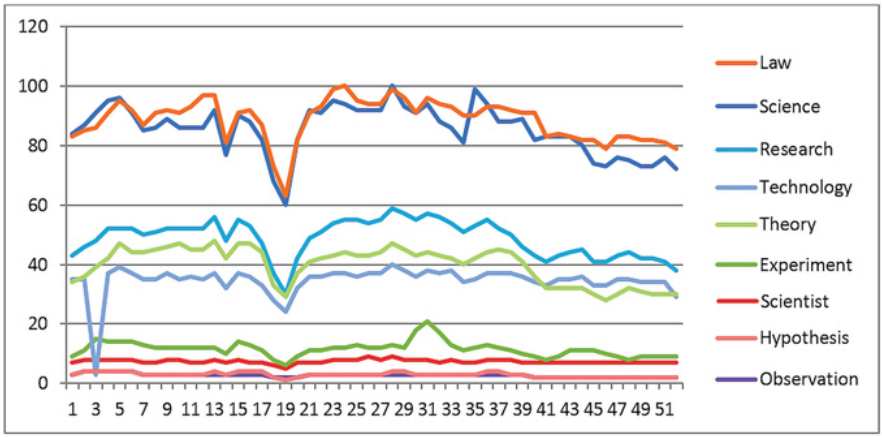
Source: Taşdere, Özsevgeç, and Turkmen (2014).

Now, we proceed to reproduce Taşdere, Özsevgeç, and Turkmen's study using GT as a Big Data application. In a sense, we will be now assessing the Nature of Science (NoS) conceptions of the *collective intelligence* (LÉVY, 1994, 1995, 1997), in which it

refers to the capacity of networked ICTs to expand the extent of human interactions and enhance the collective pool of social knowledge.

The nine concepts selected by Taşdere, Özsevgeç, and Turkmen were searched in GT across the globe during the last 52 weeks, after being translated into English to include the increased volume of worldwide data outside Turkey. The resulting graph from GT is shown in Figure 11, which compares the search frequencies relative to each other and to the time axis of these nine concepts.

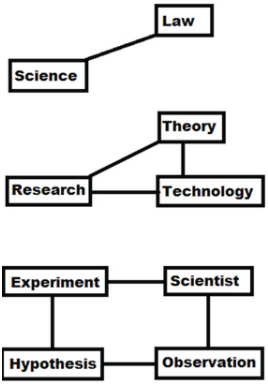
FİGURE 11 – Frequency of search of the nine keywords examined over time.



Source: Google Trends (www.google.com/trends).

When the frequency of searches is examined (Figure 11), one observes that the curves in Figure 11, corresponding to the nine concepts, arrange themselves in three layers: at the top is the layer of 'Law' and 'Science'; the second, containing 'Research,' 'Technology,' and 'Theory' comes beneath it, while at the bottom is the one with the words 'Observation,' 'Hypothesis,' 'Scientist,' and 'Experiment.' Figure 12 shows the conceptual network built from these relations.

FİGURE 12 – Conceptual network built from the curves generated by GT corresponding to the nine concepts.



Source: This research.

One may also notice that, contrary to the ones in the lowest layer, which are words that are more commonly used in more specialized contexts, and in which fewer groups are interested, the two words in the topmost layer, 'Science' and 'Law,' (Figure 12) are words with more general use in people's lives and possess 'secondary meanings', that is other meanings than those related to scientific research, but which nonetheless contribute to GT search results.

Looking at the frequency of the nine keywords surveyed according to time (Figure 11), one observes that the amount of searches increased at certain times and decreased in other times, e.g. on the week 19. This drop in all searches immediately raises the question "was any important event happening then?" If one considers that it corresponds to the week beginning on December 25, 2016, including the multinational feast days of Christmas and New Year's Eve holidays, it becomes understandable why science-related searches have dropped significantly here.

In the sequence, these nine concepts were then paired, and their two-way correlations between the frequencies of searches were calculated and shown in Table 1.

TABLE 1 – Binary correlations of the analysed words.

	Scientist	Experiment	Observation	Research	Law	Technology	Hypothesis	Theory
Science	0,8*	0,7	0,7	0,9*	0,9*	0,4	0,8*	0,8*
Scientist		0,6	0,6	0,8*	0,7	0,3	0,7	0,6
Experiment			0,6	0,8*	0,7	0,2	0,6	0,6
Observation				0,7	0,6	0	0,9*	0,7
Research					0,9*	0,5	0,8*	0,9*
Law						0,5	0,7	0,8*
Technology							0,2	0,3
Hypothesis								0,8*

* 0.8 and above were considered related and selected for building the conceptual network.

Source: This research.

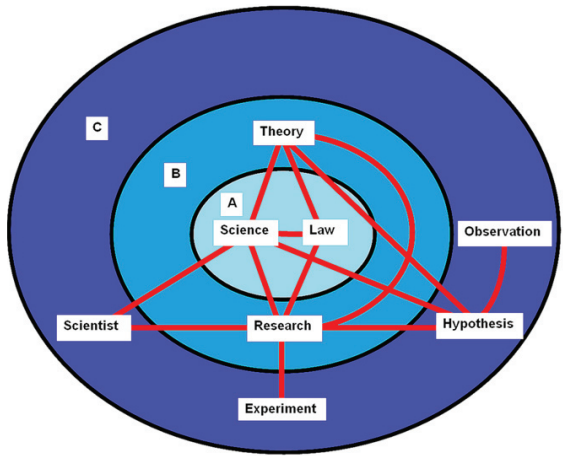
Following Taşdere, Özsevgeç, and Turkmen's procedure, those word pairs from Table 1 are then used to build a conceptual network. Only the word pairs correlated by ρ (Pearson Correlation) values of 0.8 and higher are considered related, and a (two-tailed) significance of 0.00[3] was also found for all these highly correlated pairs.

In the underlying study of Taşdere, Özsevgeç, and Turkmen (2014), there is a correlation between 'Science' and 'Technology' (Figure 10), but no such high correlation among these words was found in our Big-Data-based study. Likewise, the correlation between 'Observation,' 'Hypothesis,' and 'Experiment' observed in Figure 10 was also not found here. This result does not mean disagreement, but rather comes from a crucial difference between the samples: in Taşdere et al., it was a small sample of 23 educated training teachers, while here it was a huge number of general Google users.

When these highly related word pairs are examined, one notices that the word 'Science' has five associations, and 'Research' has six associations, and we could justifiably centre our network on the word 'Research', diverging from Taşdere, Özsevgeç, and Turkmen's one that was centred on the word 'Science'. However, since we can consider that basic research is aiming at the Nature of Science, and we intend to compare the structure of our results with the underlying research, we decided for a science-centred model for our conceptual network built from Big Data research.

Combining the previous results, we built the hierarchical network of Figure 13. The positioning of the concepts in the (A), (B), and (C) levels come from their arranging in the three layers in Figure 11. Their connections (excluding the word 'Technology', as it correlated to none of the others), come from the binary correlations of Table 1.

FİGURE 13 – Hierarchical network of binary correlations of the eight terms.



Source: This research.

[3] These correlation and significance calculations were made with the GNU PSPP program.

Three main levels emerge in our hierarchical conceptual network (Figure 13). The first level (A) is centred on 'Science' and 'Law'; the next level (B) includes the words 'Theory,' and 'Research;' and the final level (C) includes 'Scientist,' 'Experiment,' 'Hypothesis,' and 'Observation.'

As a result of this structuring, a *storytelling* (KUMAR et al., 2008) emerges: Science is seen, in a more publicly understandable level (A), as associated to 'laws', followed by a less-visible level (B) of research being associated to 'building theories', and then, in a even lesser understanding level (C), the scientists doing experiments to test hypotheses, which are confirmed or not by observation – an image of scientists' work shaped in a large degree by popular media (TAN; JOCZ; ZHAI, 2017). Nevertheless, this story may also be interpreted as a *crowdledge* (DOS SANTOS, 2015), as it is an knowledge that emerged from Big Data analysis of collective individuals' web searches.

CONCLUSION

These conclusions derived from the use of GT derive from the millions of searches made on Google and show that not only trends and distributions but also correlational studies can be done with GT.

Choi and Varian (2012) point out that the work they do on Economy using GT is based on instant data and that new searches can permanently change search results. Like our brains, whose structure changes as we learn, the structure of the stored Big Data changes as new data entries. The relations obtained are, in a sense, the average of the relations residing in the minds of existing living people. However, the network of relationships that people have in mind is not a network of real relationships because not all searchers are experts.

Tulasi (2013) draws attention to the areas of application of Big Data in Higher Education. Among others, this study shows that research on the organisation of campus planning and research centres can be done considering the relationship between science branches. For example, which department should be established or moved closer to the Department of Physics? Finding the answer to this question and planning accordingly could strengthen the scientific work at the universities.

McCosker and Wilken (2014) indicate how one of the most critical parts of Big Data studies is how the data is visualised, and the determinations of the elements that influence this visualisation. It is about how one interprets and/or understands how to visualise the data. Therefore, Big Data workers should have experience with different Big Data programs and visualisation tools. In this sense, it is possible to prepare a three-dimensional image by comparing the available data with the search frequencies and to make a more comprehensive examination by multiplying the word numbers. This detailed study would naturally have an impact on the nature of science.

Baram-Tsabari and Segev (2009b) researched the concepts of science and pseudoscience in their work through GT. The related study (BARAM-TSABARI; SEGEV,

2009a) used the then available Google's Zeitgeist ("the spirit of time") to verify the results. Not being used in the mentioned studies, Google Correlate is another Big Data application that finds the set of individual search queries whose spatial or temporal patterns are most highly correlated with the user's input (MOHEBBI et al., 2011). This approach of relating the meanings obtained by Big Data with another data source is similar to our study. Big Data workers must use multiple sources when making meaning in their work.

Pelat et al. (2009) conducted an exemplary study on how to use GT to identify some epidemics. If nine words searched similarly to the relevant study are searched in a regional sense, it may be possible to decide where to go for PhD and/or Post-Doc studies as those regions that are investing and developing in exploring trendy regions. In this sense, regional trends become relevant in future planning.

Big Data studies are a new index shift for scientific studies (DOS SANTOS, 2015). According to the works done by taking data from a limited number and a specific region, the studies that will be carried out by taking more worldwide data on will bring us closer to reality. Big Data studies will reveal work that can be performed more easily as an alternative to data errors in aggregate work done with small samples, and work that does not have interrelationships between years, but a significant multiplier effect can be created. Thus, academicians will prefer to collect data from all over the world instead of working with prospective teachers in the same building.

Big Data studies always give the impression that digital data can be analysed in a single way, but it is always important how human creativity approaches the data (MAYER-SCHÖNBERGER; CUKIER, 2013). Computers learn and develop the methods introduced to them, but when it comes to "making meaning," unusual approaches are still the work of the human brain. In this study, we proposed creative approaches to use Big Data in Education.

REFERENCES

ALPHABET INC. *Alphabet Announces Second Quarter 2017 Results*. Mountain View, CA: Alphabet Inc., 24 jul. 2017. Disponível em: <https://abc.xyz/investor/news/earnings/2017/Q2_alphabet_earnings/>. Acesso em: 4 set. 2017.

BANTIMAROUDIS, Philemon. A mediated assessment of Samuel Huntington's "Clash of Civilizations": The cultural framing hypothesis. *International Journal of Media & Cultural Politics*, v.11, n.1, p.73–85, 1 mar. 2015. DOI: 10.1386/macp.11.1.73_1.

BARAM-TSABARI, Ayelet; SEGEV, Elad. Exploring new web-based tools to identify public interest in science. *Public Understanding of Science*, v.20, n.1, p.130–143, 9 out. 2009a. DOI: 10.1177/0963662509346496.

BARAM-TSABARI, Ayelet; SEGEV, Elad. Just Google it! Exploring New Web-based Tools for Identifying Public Interest in Science and Pseudoscience. 2009b, Raanana: The Open University of Israel, 2009. p.20–28. DOI: 10.1177/0963662509346496.

CAVAZOS-REHG, Patty et al. Monitoring marijuana use and risk perceptions with Google Trends data. *Drug and Alcohol Dependence*, v.146, p.e242–e243, 1 jan. 2015. DOI: 10.1016/j.drugalcdep.2014.09.126.
CHOI, Hyunyoung; VARIAN, Hal R. Predicting the Present with Google Trends. *Economic Record*, v.88, p.2–9, 27 jun. 2012. DOI: 10.1111/j.1475-4932.2012.00809.x.
DOS SANTOS, Renato P. Are our students really interested in Science? Or does Google Trends show a socially desirability bias in Brazilian public opinion surveys? *Acta Scientiae*, v.18, n.2, p.531–549, May 2016.
DOS SANTOS, Renato P. Big Data: Philosophy, Emergence, Crowdledge, and Science Education. *Themes in Science and Technology Education*, v.8, n.2 – Special Issue on Big Data in Education, p.115–127, Dec. 2015.
DOSTÁL, Jiří. The definition of the term "Inquiry-based instruction". *International Journal of Instruction*, v.8, n.2, p.69–82, 30 Jul. 2015.
GOOGLE INC. *How Trends data is adjusted*. Disponível em: <https://support.google.com/trends/answer/4365533?hl=en-GB>. Acesso em: 11 maio 2016a.
GOOGLE INC. *Trends graphs and forecasts*. Disponível em: <https://support.google.com/trends/answer/4355164?hl=en-GB>. Acesso em: 29 abr. 2016b.
GOOGLE INC. *Where Trends data comes from*. Disponível em: <https://support.google.com/trends/answer/4355213?hl=en-GB>. Acesso em: 11 maio 2016c.
GUO, Shesen; ZHANG, Ganzhou; ZHAI, Run. A potential way of enquiry into human curiosity. *British Journal of Educational Technology*, v.41, n.3, p.E48–E52, maio 2010. DOI: 10.1111/j.1467-8535.2009.00949.x.
HEIBERGER, Raphael H. Collective Attention and Stock Prices: Evidence from Google Trends Data on Standard and Poor's 100. *PLoS ONE*, v.10, n.8, p.e0135311, 10 ago. 2015. DOI: 10.1371/journal.pone.0135311.
KUMAR, D. et al. Algorithms for Storytelling. *IEEE Transactions on Knowledge and Data Engineering*, v.20, n.6, p.736–751, jun. 2008. DOI: 10.1109/TKDE.2008.32.
LÉVY, Pierre. Collective Intelligence and Its Objects: Many-to-Many Communication in a `Meaning World'. *Doors of Perception 3*. [S.l: s.n.], 1995. Disponível em: <http://museum.doorsofperception.com/doors3/transcripts/Levy.html>.
LÉVY, Pierre. *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Tradução Robert Bononno. Cambridge, MA: Plenum Trade, 1997. v.1.
LÉVY, Pierre. *L'intelligence collective – Pour une anthropologie du cyberspace*. Paris: La Découverte, 1994.
MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: Hodder, 2013.
McCOSKER, Anthony; WILKEN, Rowan. Rethinking "big data" as visual knowledge: The sublime and the diagrammatic in data visualisation. *Visual Studies*, v.29, n.2, p.155–164, 4 maio 2014. DOI: 10.1080/1472586X.2014.887268.
MOHEBBI, Matthew H. et al. *Google Correlate Whitepaper*. Menlo Park, CA: Google Inc., 9 jun. 2011. Disponível em: <http://www.google.com/trends/correlate/whitepaper.pdf>. Acesso em: 28 abr. 2013.
MURGIA, Madhumita. How smartphones are transforming healthcare. *FT Magazine*, 12 jan. 2017. Disponível em: <https://www.ft.com/content/1efb95ba-d852-11e6-944b-e7eb37a6aa8e>. Acesso em: 29 out. 2017.

NSOESIE, Elaine O.; BROWNSTEIN, John S. Computational Approaches to Influenza Surveillance: Beyond Timeliness. *Cell Host & Microbe*, v.17, n.3, p.275–278, 11 mar. 2015. DOI: 10.1016/j.chom.2015.02.004.
PAPERT, Seymour A. What's the big idea? Toward a pedagogy of idea power. *IBM Systems Journal*, v.39, n.3.4, p.720–729, 2000.
PELAT, Camille et al. More diseases tracked by using Google Trends. *Emerging infectious Diseases*, v.15, n.8, p.1327–8, ago. 2009. DOI: 10.3201/eid1508.090299.
PENCE, Harry E.; WILLIAMS, Antony J. Big Data and Chemical Education. *Journal of Chemical Education*, v.93, n.3, p.504–508, 8 mar. 2016. DOI: 10.1021/acs.jchemed.5b00524.
PROVOST, Foster; FAWCETT, Tom. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, v.1, n.1, p.51–59, mar. 2013. DOI: 10.1089/big.2013.1508.
SCHEITLE, Christopher P. Google's Insights for Search: A Note Evaluating the Use of Search Engine Data in Social Research. *Social Science Quarterly*, v.92, n.1, p.285–295, mar. 2011. DOI: 10.1111/j.1540-6237.2011.00768.x.
SCOTT, Steven L.; VARIAN, Hal R. Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, v.5, n.1/2, p.4, 2014. DOI: 10.1504/IJMMNO.2014.059942.
SINCLAIR, Betsy; WRAY, Michael. Googling the Top Two: Information Search in California's Top Two Primary. *California Journal of Politics and Policy*, v.7, n.1, p.6, 5 fev. 2015. DOI: 10.5070/P2CJPP7125443.
TAN, Aik-Ling; JOCZ, Jennifer Ann; ZHAI, Junqing. Spiderman and science: How students' perceptions of scientists are shaped by popular media. *Public Understanding of Science*, v.26, n.5, p.520–530, 18 jul. 2017. DOI: 10.1177/0963662515615086.
TAŞDERE, Ahmet; ÖZSEVGEÇ, Tuncay; TÜRKMEN, Lütfullah. Bilimin Doğasına Yönelik Tamamlayıcı Bir Ölçme Aracı: Kelime İlişkilendirme Testi (A Complementary Measurement Tool for the Nature of Science: Vocabulary Association Test). *Fen Bilimleri Öğretimi Dergisi (Journal of Science Teaching)*, v.2, n.2, p.129–144, dez. 2014. DOI: 10.12973/eurasia.2015.1367a.
TULASI, Bomatpalli. Significance of Big Data and Analytics in Higher Education. *International Journal of Computer Applications*, v.68, n.14, p.21–23, 18 abr. 2013. DOI: 10.5120/11648-7142.
YIN, Chengjiu et al. Learning by Searching: A Learning Environment that Provides Searching and Analysis Facilities for Supporting Trend Analysis Activities. *Journal of Educational Technology & Society*, v.16, n.3, p.286–300, jul. 2013. Disponível em: <http://www.ifets.info/others/abstract.php?art_id=1394>. Acesso em: 7 fev. 2014.
ZHANG, Meilan et al. Tracking the Rise of Web Information Needs for Mobile Education and an Emerging Trend of Digital Divide. *Computers in the Schools: Interdisciplinary Journal of Practice, Theory, and Applied Research*, v.32, n.2, p.83–104, 2015. DOI: 07380569.2015.1030531.