MDPI

Article

Revealing trends in geophysics using metadata and text analysis

Timofey Eltsov [1,†], Maxim Yutkin [2], Tadeusz W. Patzek [3]

[1] Ali I. Al-Naimi Petroleum Engineering Research Center, King Abdullah University of Science and Technology; timofey.eltsov@kaust.edu.sa
[2] Ali I. Al-Naimi Petroleum Engineering Research Center, King Abdullah University of Science and Technology; maxim.yutkin@kaust.edu.sa
[3] Ali I. Al-Naimi Petroleum Engineering Research Center, King Abdullah University of Science and Technology; tadeusz.patzek@kaust.edu.sa
* Correspondence: timofey.eltsov@kaust.edu.sa; Tel.: +966128087182
† Current address: 4700 KAUST, Thuwal, 23955-6900, Saudi Arabia

Version February 18, 2020 submitted to Geosciences

Abstract: Professional language evolution reveals the development of geophysics: researchers enthusiastically describe new methods of survey, data processing techniques, and objects of their study. Geophysicists publish their cutting-edge research at international conferences proceedings to share their achievements with the world. Tracking changes in the language allows one to identify trends and the current state of the science. Here, we describe the text and metadata analysis of the last 38 Annual Conferences organized b y the Society of Exploration Geophysicists, one of the biggest geophysical gatherings. We split 24,500 articles into words and phrases and analyze the change in their usage frequency over time. We find that in 2019 the phrase "neural network" is used more often than "field data." The word "shale" has become less commonly used, but the term "unconventional" is growing in occurrence. The number of publications from oil companies reflects their financial situation; the number of papers from the academia of various countries indicates government financing of research. The USA academia has the most significant number of publications; in 2019, the number of papers from China was almost equal to those of the USA. An analysis of conference materials and metadata allows one to identify trends in a specific field of knowledge and predict the development in the near future.

Keywords: geophysics; web data analysis; data mining; data analysis; text mining; words analysis;

1. Introduction

The last four decades showed a tremendous change in geophysics. An increase in computing power and technological progress allowed geophysicists to solve more and more complicated tasks. At the same time, the field of application of geophysics is expanding; the market of geophysical services is changing. We assume that a change of geophysical tasks, applications, geography, and technology will inevitably lead to a shift in the professional language. If one can track changes in the frequency of terms used in recent years, one can shed light on the current state of the industry and possibly predict future changes. Authors apply language processing methods to analyze changes in the professional language in geophysics.

The biases of different origin complicate big data [1]. In machine learning, the difference between training data set and test data set can cause biases. Massive sample study can lead to bias associated with an error resulting from sampling or study design [2]. Supposedly, it is better to have a smaller and more representative data set rather than a much bigger but biased data. We want to understand

what the modern geophysical language looks like and what the future of geophysics will be. In this paper, we analyze only scientific articles presented at the Society of Exploration Geophysicists (SEG) Annual Conference and Exhibition. The committee selects the papers for the conference each year; this is the initial filtering. Also, it is worth noting that presenting at such a meeting is a demonstration of the technical capabilities of industrial companies and the scientific viability of academic institutions. Each annual conference proceedings is a cross-section of the state of geophysics, and we use it for analysis and predictions.

The SEG Annual Conference and Exhibition is one of the biggest gatherings of geophysicists in the world. Abstracts of the SEG Annual Conferences are a representation of the state of geophysical science, approximated mainly to the oil and gas industry. Articles in the electronic version for the 38 years are available for analysis [3]. The SEG conducted all their Annual conferences in the USA, and the last one was in San Antonio, TX. For analysis, the authors selected the proceedings of the SEG Annual Conference, as the most representative set, that reflects state-of-art-technologies in geophysics. Each conference proceedings is a reflection of the state of the industry in a particular year since, at this event, both academic institutions and the industry present their best achievements in the field.

Besides conference proceedings, one can use journal articles for data mining as the volume of the data for one year is comparable to the SEG Annual Conference and Exhibition Proceedings. The number of publications per year is smaller, but they consist of full-size papers. However, the release of articles in journals is carried out periodically, e.g., monthly or quarterly; at the conference, this happens once a year. The research materials are usually published in journals and reported at conferences; the proceedings include many of the results from full-sized articles. Moreover, the number of research teams presenting their work is several times larger in the case of analysis of conference materials compared to the study of one particular journal. SEG Annual Conference proceedings represent a collection of scientific research from a large number of scientific teams in one place for each of the 38 years. This approach allows one to conduct a unique study and trace the dynamics of changes in the industry.

2. Materials and Methods

We used the OnePetro online library [4] to get metadata of the reviewed papers. The OnePetro website offers ample opportunities for analyzing metadata. Along with OnePetro, CrossRef service can be used for metadata analysis. Different spelling versions and typos affected the study of affiliation. Besides, many organizations have since ceased to exist (acquired, bankrupt, split, *etc.*), which also complicated the analysis. We use open-source Python libraries: to transform, filter and process the text, and get metadata: TextBlob, NLTK (Natural Language Toolkit), argparse, Pandas, Scrapy, Requests-HTML, sqlite3, and NumPy. For printing the graphs, we use Matplotlib, Plotly, PIL (Python Imaging Library), and others.

For text analysis, we used digital versions of the SEG Annual Conference proceedings that have been available online for 38 years. Fig. 1 shows the workflow scheme. Abstracts from the 1980s consisted of 1 or 2 pages; in the 1990s, it increased to four pages per abstract. We digitized articles in PDF format from the SEG digital library website, converted them into plain TXT format using "pdftotext" with "nopgbrk" (ignore page breaks), "enc ASCII7" (sets ASCII7 encoding for the output) and "eol" (sets the end-of-line convention) flags. The text damp was filtered to remove common words, misspellings, *etc.* from a NLTK dictionary, "stopwords." After the initial filtering, we tokenized the text by year and obtained si-, bi-, and trigrams[1]. Further, we counted the number of times each word or phrase was repeated. Finally, the entire text for 38 years is a three-dimensional array of words and phrases with the corresponding number of repetitions for each year. We then analyze the list of words and phrases during observation time and display the results in a graphical form.

[1] "sigram" - a word, "bigram" - two-word phrase, "trigram" - three-word phrase

While digitalizing abstracts from the 1980s, recognition errors, merged words, and typos occur. Therefore, we present metadata analysis for the entire period; however, the phrase count is done only for the period from 1990 to 2019.

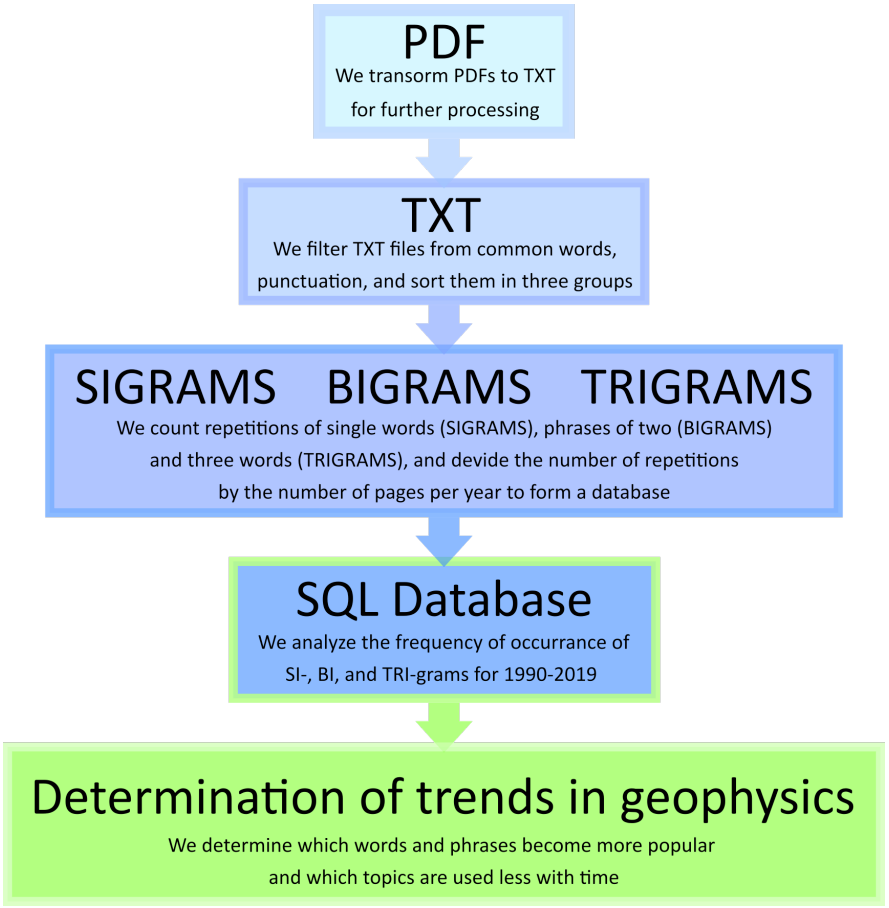

Figure 1. Data processing workflow.

In total, we analyzed 24,500 papers consisting of more than 57 million words or more than 383 million symbols. The number of non-unique authors for the entire time span chosen in this paper exceeds 75 thousand.

3. Results

3.1. Academia and Industry

More than 2400 industry companies and academic institutions from eighty-six countries heve presented their research at the SEG Annual Conference so far. The five companies with the most significant number of publications in the SEG Annual Conference are 1) Schlumberger, 2) WesternGeco, 3) CGG (Compagnie Générale de Géophysique), 4) BGP Inc. (BGP Inc., China National Petroleum Corporation), and 5) BP plc (formerly The British Petroleum Company plc and BP Amoco plc). These five companies accounted for about 30% of all affiliations in the past ten years. Schlumberger itself constitutes about 8% of all affiliations, with WesternGeco adding about 5% in the past ten years. The five most highly represented universities in SEG Annual Conferences over the 38 years are 1) Colorado School of Mines, 2) University of Houston, 3) China University of Petroleum, 4) Stanford University, and 5) Delft University of Technology.

Fig. 2 reveals the average number of papers for the academia by country. Each country is represented by a unique color; the size of the circles is the average number of publications. The inset

in the lower-left corner shows a Europe zoom-in. The most considerable contribution is from the universities in the United States of America (the USA, see the five above) followed by universities in China (China University of Petroleum, Jilin University, and Tongji University), and Canada (University of British Columbia). The rest is shared among the Netherlands, France, Germany, and Russia.

China holds the leading position in Asia, followed by smaller but notable contributions from South Korea, Japan, and India. In South America, Brazil academia publishes the most abstracts. In the Middle East, the most represented country at the SEG Annual Meetings is the Kingdom of Saudi Arabia over the past ten years. In 2009 the average number of publications was about one, and in 2019, it is more than 23, which is indeed impressive. None of the academia of the other countries have shown such rapid relative growth in recent years. The total number of publications from the top-50 countries is presented in the appendix, and the full list can be accessed here [5].

The circle on Antarctica represents incomplete affiliations or affiliations with typos that were not correctly recognized; therefore, it was impossible to determine the location. It provides an estimate of the total error of the analysis.

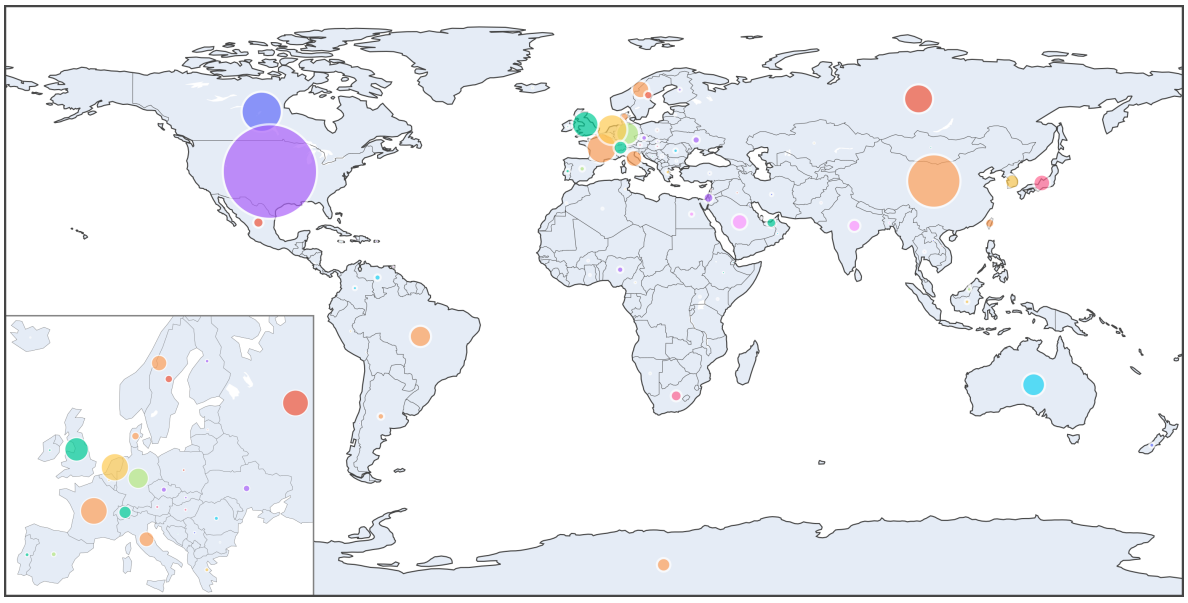

Figure 2. The total number of academic publications by country. Europe is in the zoomed inset on the lower left. The circle on Antarctica represents erroneous affiliations and serves as an error indicator.

Fig. 3 compares the annual number of publications from industry and academia. Both contributors show a steady growth over the years, which is associated with an increase in fossil fuel consumption, oil price, and constant-growth-economic paradigm. However, on the finer scale, there is a weak correlation with the oil price change. For example, the number of academic publications was hardly affected by the two recent crises in 2008 and 2014. On average, the number of industrial publications is only partially influenced by the oil price dynamics resulting in a slower growth rate in the last few years. It appears that, on average, the industry became more efficient in research expenditure optimization, which enabled them to maintain a high number of publications and even a slow but consistent growth during the shrinking market.

Curiously, the number of publications from the academia in the last year has fallen significantly. Fig. 3 indicates a decrease in the number of publications from the USA academia in 2019 compared to 2018. Perhaps this was due to a reduction of state funding of higher educational institutions [6]; see blue curve drop in 2018-2019 Fig. 5.

Fig. 4 shows that the number of co-authors per paper has increased and we observe a correlation with the world trend exemplified by a related field of the Earth and Planetary sciences. The increase in the number of authors per paper is a worldwide trend [7]; scientific research is becoming more

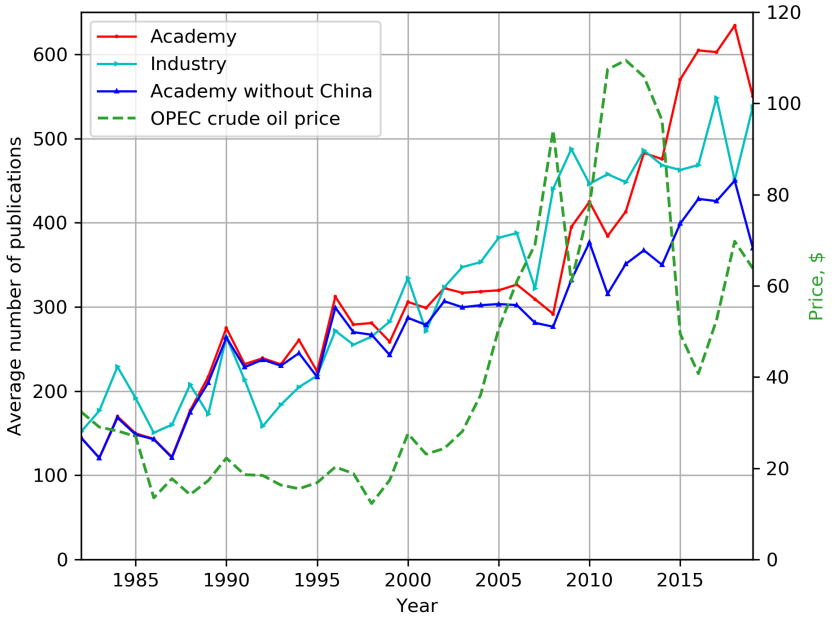

Figure 3. The annual breakdown of net industry and academia publications.

interdisciplinary and thus more collaborative. The SEG average co-author number almost flattens out at 3.6 co-authors per paper, but in 2019 the number of authors per paper increased, reaching 3.9. With that, we see an increase in the number of organizations involved in the SEG Annual Conference.

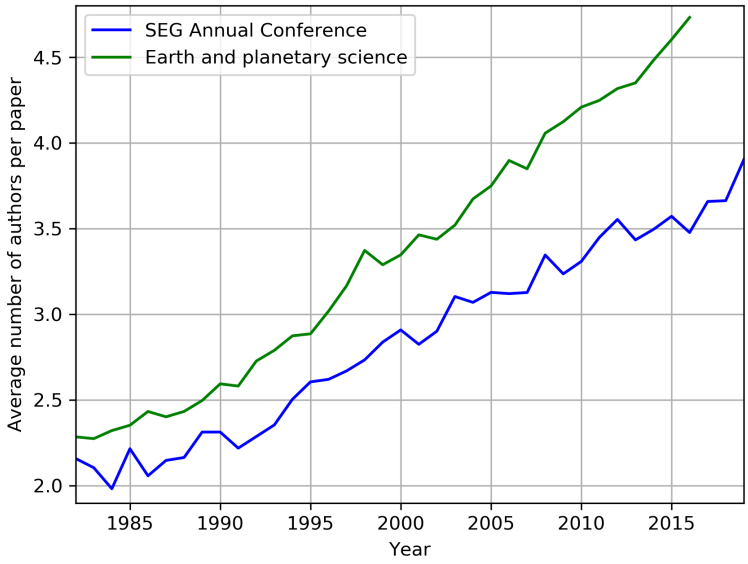

Figure 4. Number of co-authors growth rate for the SEG Annual Conference and for the Earth and Planetary science [7].

Fig. 5 shows a breakdown of academic publications by country. The USA academy is ahead of everyone in the number of papers published annually, as well as total articles published. The Netherlands and Canada, have regular contributions, and their publication activity is constant over time. Other countries, like France and Germany, seem to follow the crude oil price trend. In contrast, China maintains a steady growth rate. In 2006 the Chinese government initiated a powerful program of research development, "Medium and Long-term Plan for the Development of Science and Technology

(2006–2020)" with a target of 2.5% GERD/GDP ratio by 2020 [8]. The strong support of the government for geoscience, allowed the Chinese academia to exhibit the fastest growth between 2008 and 2015. In 2013, we observed a 15% increase on R&D spending by China compared to 2012 [9]. The number of publications by the Chinese academia is now almost equal to those of the USA.

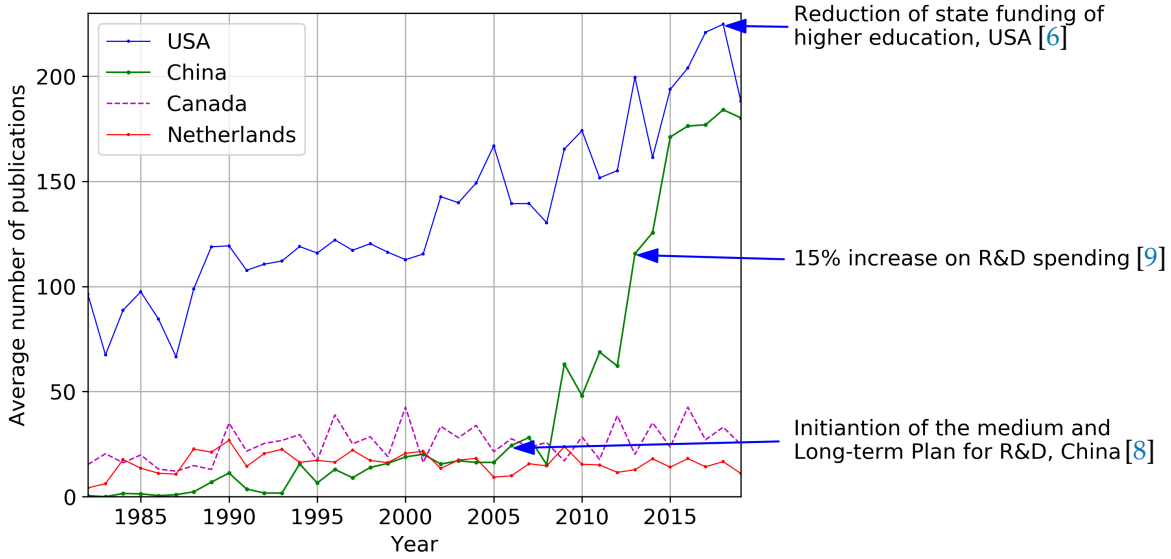

Figure 5. Average number of publications by academia of the TOP-four countries academia.

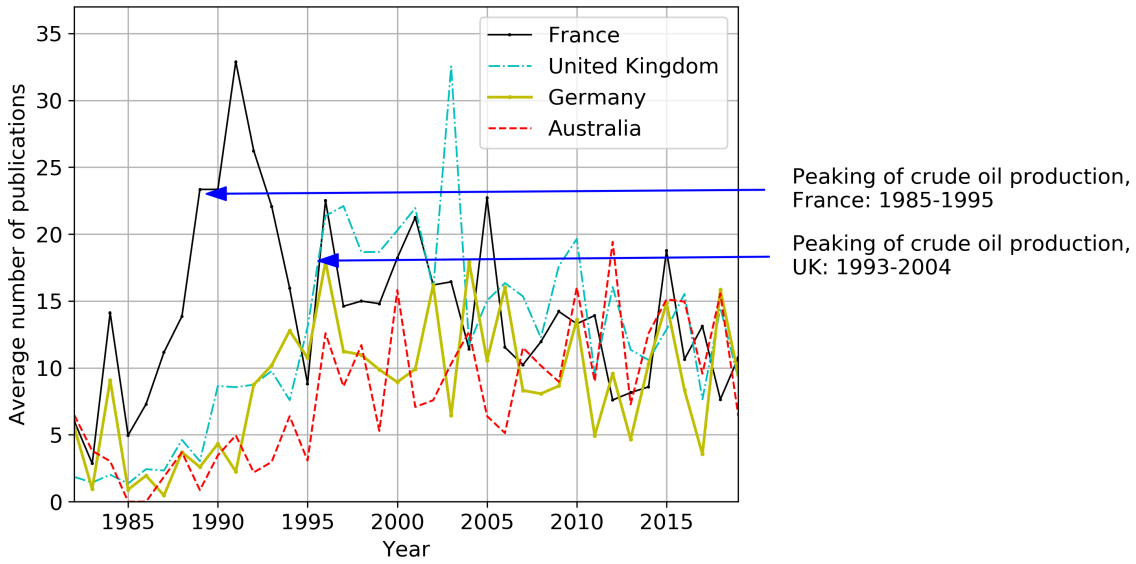

Figure 6. Average number of publications by France, UK, Germany, and Australia. Crude oil production data is presented online [10].

Next, we perform a similar analysis for industry publications. Fig. 7 shows the average number of papers by oilfield service companies. The most frequent guests at the SEG Annual Conferences are Schlumberger, WesternGeco, CGG, and BGP. Although WesternGeco is a part of Schlumberger, we show them separately according to the affiliation. Schlumberger dominates industrial geophysics research, followed by CGG and BGP. In general, the number of publications by the major oilfield service companies grew steadily. Although oilfield service providers are dependent on oil prices, we surprisingly observe that after the 2014 crisis, the number of Schlumberger publications peaked for several consecutive years, followed by a decline in 2018. It should be noted that in 2014 Schlumberger

reported an outstanding revenue of $48.6 billion. The dynamics of Schlumberger's papers reflect the dynamics of oil prices with a few years offset. The number of CGG publications number follows crude oil prices too, but since 2014, the number of publications from CGG has only declined. CGG Annual report states extremely difficult market environment [11], and cost reduction measures: reduction in the number of employees from 11060 to 7353, 55% general and administrative cost-cutting, 64% cost of marine monthly structure cost-cutting[2]. In January 2020 CGG reported its exit from marine acquisition, sale of ships and measuring equipment to Shearwater company [12]. This news suggests that we will see fewer publications from CGG in the coming years.

The change in the number of BGP publications shows a similar trend with the crude oil prices with one or two years of delay. In 2019 BGP became a leader by the number of publications among oil service providers.

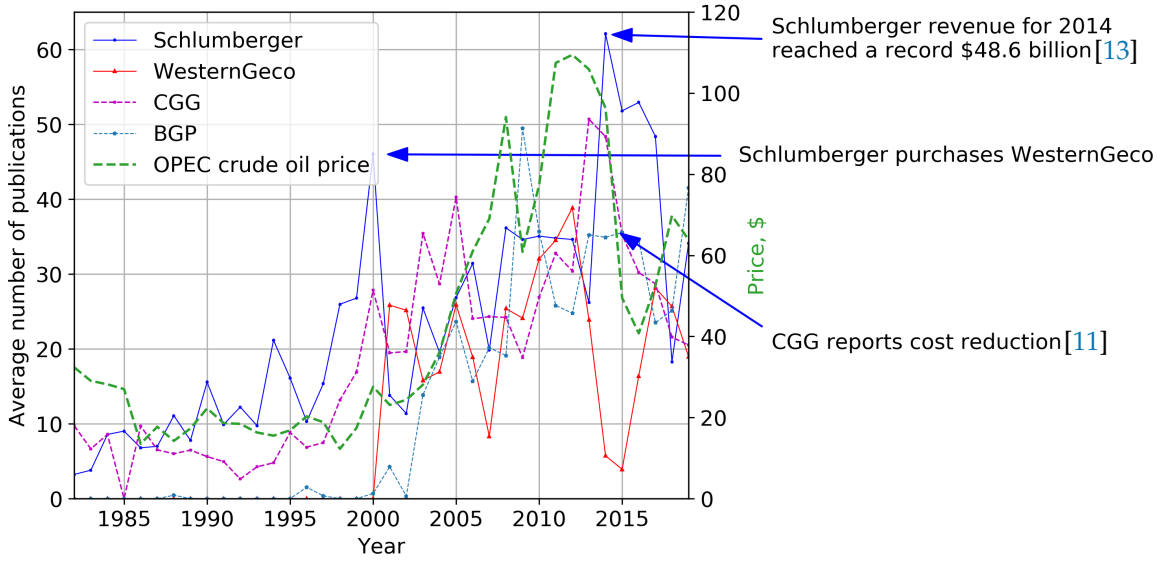

Figure 7. The average number of publications by oil-service companies and OPEC crude oil prices.

Many oil and gas companies that no longer exist made significant contributions to the SEG Annual Conference in the 1980s and early 1990s. They are Arco Oil and Gas Co., Mobil E&P, OYO Corporation, Statoil, and others. These companies either merged with others, changed their names, or were acquired.

Fig. 8 displays five oil production companies with the most significant number of publications. The picture is conceptually different from oil service companies. For example, the number of publications by BP and ExxonMobil peaked in 2005 and nowadays, it is declining. 2005 was an outstanding year for ExxonMobil, with a net income of 36 billion and a 31% increase in the number of employees [15]. We observe steady growth in many economic indicators of the company since the beginning of 2000. At the same time, a decrease in the number of publications indicates a difficult period for the company. For instance, in 2014, we found only one paper from ExxonMobil, which has not happened over the past 15 years. The 2014 ExxonMobil Summary Annual Report [16] shows that compared to 2013, market valuation at the end of the year decreased by 12%, and we observe the decline in the stock market price of ExxonMobil in 2015. The decrease in profits immediately affects research financing. Saudi Aramco demonstrates steady growth; it had the biggest number of publications of all production companies in 2017 and 2018. Interestingly enough, the leadership was taken by PetroChina in 2019, followed by Shell and Saudi Aramco.

[2] Comparing between 2013 and 2015

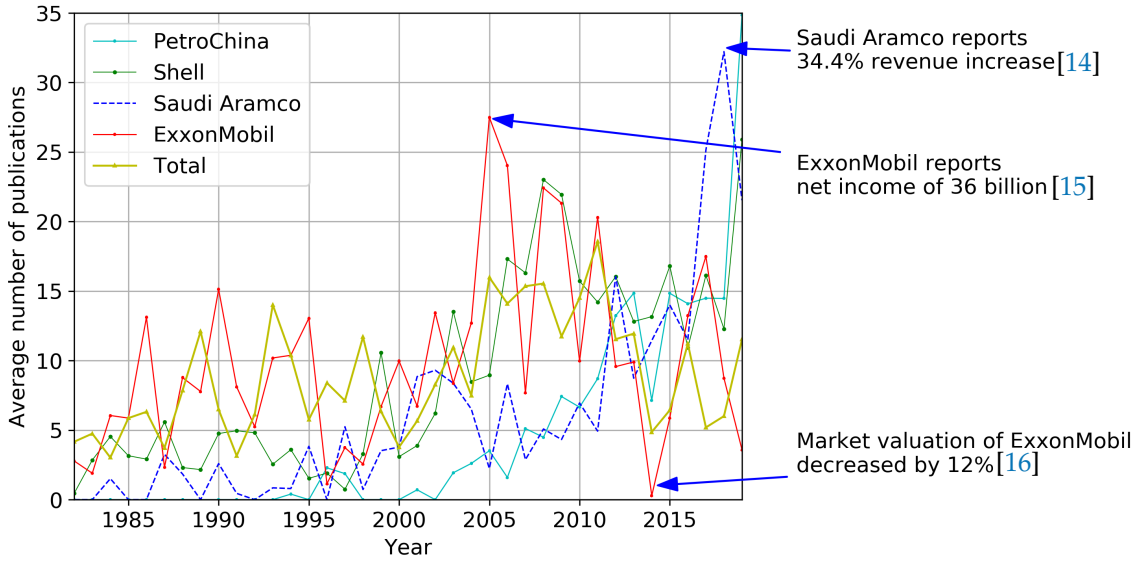

Figure 8. The average number of publications by oil production companies.

3.2. Technical terms analysis

In this section, we present our analysis of manuscript texts from the SEG Annual Conference. However, instead of focusing on average text length or sentence complexity, we look into the technical side. For example, we analyze and compare the frequency of occurrence of technical terms, such as "data" or "velocity." Such an analysis sheds some light on technology development and trends in the field.

The usage of the most frequently used English words ("the," "of," "and," "to," *etc.*) per page remains unchanged over the last decades. Therefore, we normalize the data to the number of pages of all articles every year. Often pages are not entirely filled with text; there are many graphs and formulas. Moreover, we know precisely the number of characters used, and we can estimate the number of pages. We calculate the average number of pages, Np, for each year using the formula: $Np_i = \frac{Ns_i}{3000}$, where i is the corresponding year, Ns - number of symbols, 3000 is the number of characters for the common one spaced web page. The estimated number of analyzed pages is 127.9 thousand. When analyzing the graphs in the present paper, one can state the number of times the phrase occurred per page each year.

3.2.1. Most common words and phrases

Fig. 9 shows the most commonly used words, two- and three-word phrases that appeared in conference materials from 1990 to 2019. The most frequent words are "data," "model," "velocity," "seismic." Through the whole period of the study, the word "data" was mentioned more than 377700 times, "seismic" 252400, "model" more than 251500 times, and "velocity" more than 223300 times in thirty-eight years. For comparison, the mention of the word "that" was 324250 times. Most of the three- and two-word phrases are devoted to seismic exploration and seismic data processing.

Frequent use of these words tells us that most of the articles are about seismic exploration and seismic data processing. The terms "wellbore" and "logging" were more popular during the 1980s, and now their relative occurrence is declining.

While net average values provide information regarding the key concepts used over time, they are of less interest for absolutely the same reason. A more exciting approach is to monitor the evolution of other technical terms that constitute a subfield in geoscience or pertain to other disciplines. Such an analysis, however, is infinite. We limited the scope of this discussion to the objectives of the study, methods of data gathering and processing, shales, and neural networks. We also considered the fastest growing and declining trends in the publications.

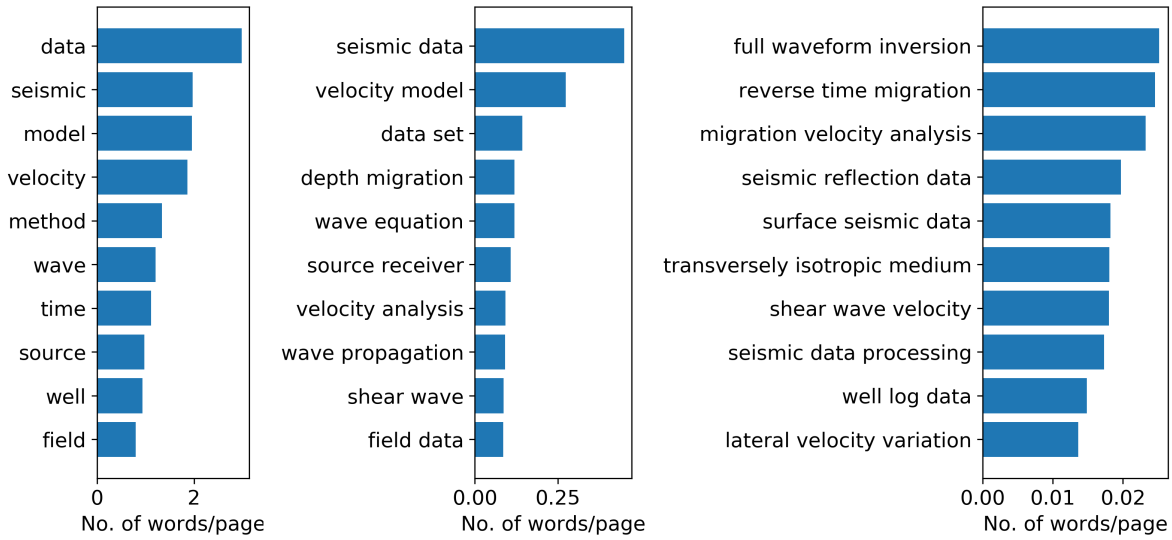

Figure 9. The average frequency of single words (left), two-word phrases (middle), and three-word phrases (right) per page for the most frequently used terms (the 1990s-2019). The total number of pages is 127.9 thousand.

3.2.2. Objects of the study

Fig. 10 breaks down the most used types of rocks. Each of the words on the left includes the most common names of rocks, e.g., sedimentary: shale, sandstone, conglomerate, carbonate, etc.; igneous: granite, diorite, basalt *etc.*; metamorphic: gneiss, phyllite, slate, *etc.* It shows the relative distribution of the objects of study: the majority of research deals with the sedimentary rocks. Terms that describe igneous rocks are used about ten times less than sedimentary, and the least used are metamorphic rocks related terms. The right part of Fig. 10 shows the occurrence of rock types with time. The shale revolution starting in 2007 is clearly notable. The most occurring names of rocks are "shale," "sandstone" and "carbonate." We note how "shale" peaks around 2015 and starts declining afterwards. Besides, there is a steady increase in the appearance of "carbonate" (the 1990s - 2005), while "sandstone" is used evenly over the years. An attentive reader will notice that during the growth of the use of the word "shale," the fluctuation of the use of the words "sandstone," and "carbonate" decreased.

We breakdown the sum of the names of geophysical methods used from 1990 to 2019, Fig. 11, left. They practically do not change over time, and we show the total in a pie chart. The whole pie chart is the sum of all the words we use in Fig. 11[3]. We show the occurrence of the most common geophysical methods, and it gives us an estimate of SEG Annual Conferences content. Three-fourths of the material relates to the collection and processing of seismic data; the remaining quarter accounts for all other methods. We see that the primary method discussed at the SEG Annual conference is "seismic," its usage is an order of magnitude higher than other methods and it is still growing in the frequency of occurrence. It is worth noting, that the word "seismic" is mentioned about four times more often than the word "geophysics." On the right part of Fig. 11, we breakdown the names of the main resources. We observe a slight increase in the frequency of the words "gas" and "water" from 1995 to 2015. Perhaps this is due to the increase in reservoir modeling related research. Figs. 10 and 11 show that for the last 30 years there are no significant changes in the use of geophysical methods and objects, with the exception of an increase in the frequency of occurrence of "shale" from

[3] Each of the words represents the sum of the related words: "seismic," "seismics"; "magnetic," "geomagnetic," "aeromagnetic"; "electromagnetic," "em"; "gravity," "gravimetry," "gravimetric"; "electric," "geoelectric"; "logging," "borehole geophysics."

2010 to 2018 and a slight increase in the occurrence of "water" and "gas" from 1995 to present days. Significant changes occur in the use of terms related to specific methods of geophysical survey and data processing. This will be discussed in subsequent sections of the paper.

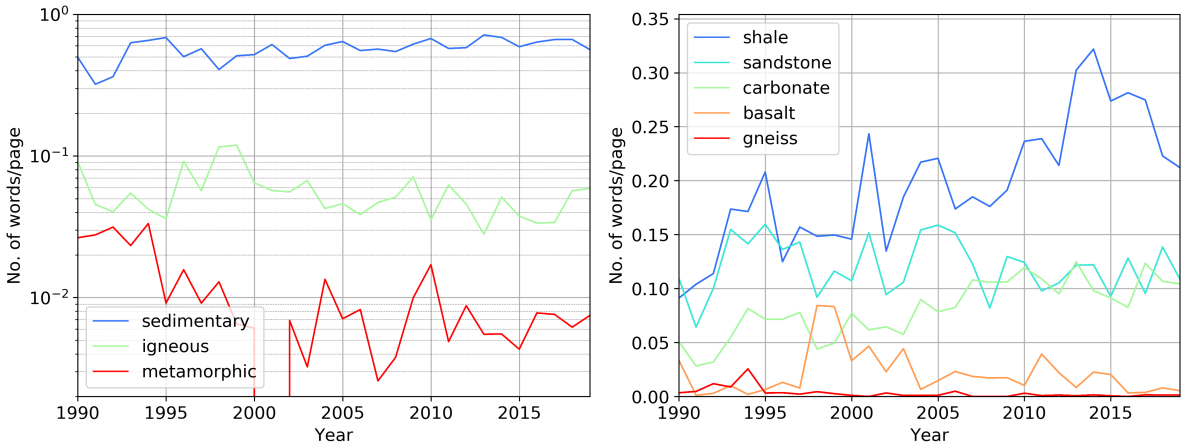

Figure 10. Frequency of use of different rock types (left) and most often used rock names (right). Rock types include most common rocks, e.g., sedimentary: shale, sandstone, carbonate; igneous: granite, diorite, basalt, *etc.*; metamorphic: gneiss, phyllite, slate, *etc.*

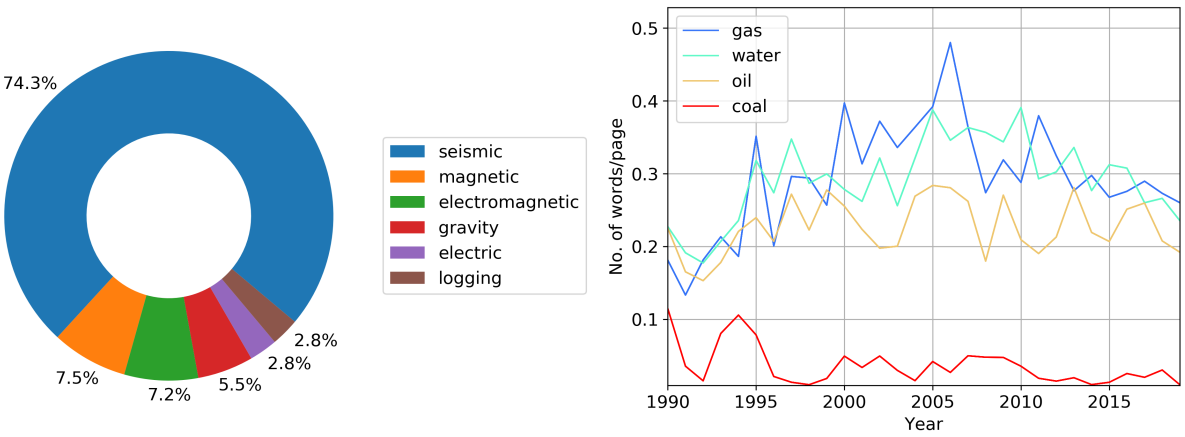

Figure 11. Geophysical methods of survey (left) and the most frequently used names of fossil fuels (right).

3.2.3. Processing and data acquisition methods

Of all the three-word phrases, the most frequently used now is "full waveform inversion" (Fig. 12), and it is still growing together with the abbreviation "FWI." The second one is "reverse time migration," the 2019 top three closes with "convolutional neural network." Fig. 12 shows how the occurrence of "prestack depth migration" was substituted by "full waveform inversion" and "reverse time migration." The frequency of occurrence is higher if we consider abbreviations. It is interesting to note that the abbreviations "FWI" and "RTM" are used more often than "PSDM" even when it was much more accessible. Perhaps this suggests a tendency to reduce and simplify the terms. The growth of the use of some terms inevitably supplants other words, provided that the volume of published material is approximately the same. While reviewing conference proceedings for the 38 years, we found many terms that were popular before, but do not find application in the modern world. Fig. 13, left, breaks down other trends in seismic data processing algorithms. We see that "machine learning" appears in the SEG Annual Conference proceedings more often in the past few years. The occurrence of

the word “broadband” started to increase in the early 2010s with a corresponding decline in 2016-2018; it began to grow again in 2019. Using a wider frequency range and inclusion of the low frequencies proved to contribute to better resolution, penetration, and inversion [17]. Besides, we see an emergence in the rate of occurrence of the “Marchenko” method. “Marchenko” is a set of data-driven methods that help us to project surface seismic data to points in the subsurface, it relates Green’s function from a virtual source inside a medium to the reflection response at the surface of that medium [18,19]. “Markov”-chain-based approach is able to account for the change in seismic response of damaged structures [20], and it correlates with the occurrence of the word “seismicity.” The term “seismicity” is used for induced seismicity risk estimation [21], mine development [22], and other applications. It is known that “machine learning” and “neural networks” have recently significantly developed towards image recognition, and the in seismic data processing, and we will devote a separate part of the paper to this issue. Fig. 13, right, shows classic methods of seismic data processing and related terms. We see the rise and decrease in the appearance of these words in the last ten years, these methods were developed in the 1990s and now they have already been studied enough, and therefore their usage is decreasing. It should be noted that despite the decrease in the number of occurrence of the words “Kirchhoff” (migration), “CMP” (Common Mid Point) gather, “NMO” (Normal Moveout), “velocity analysis,” and “interferometry,” all of them are used in industrial seismic. These words are still used quite often, but the time of research and development of methods associated with these words was the 1990s and early 2000s. The decrease in the frequency of occurrence suggests that research on this topic has decreased.

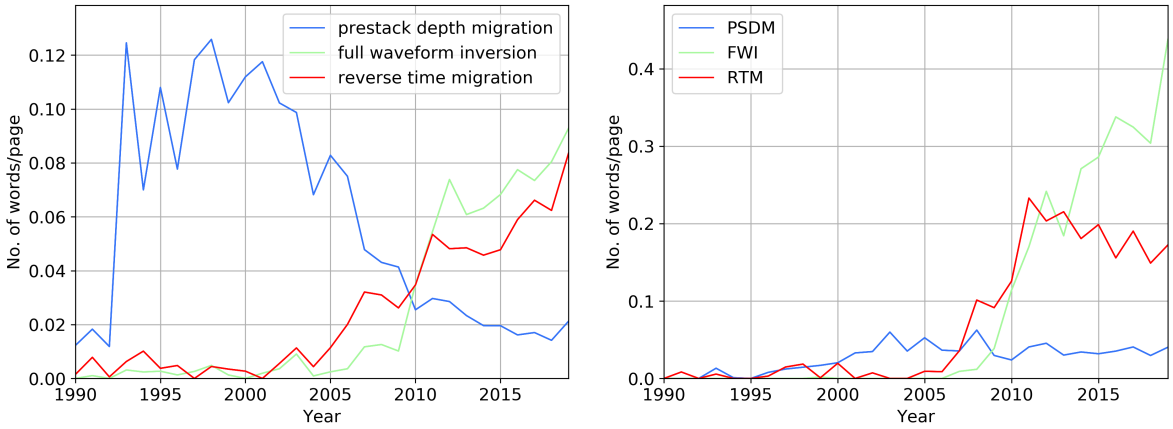

Figure 12. Change in the use of seismic data processing methods: full expression (left) and abbreviations (right).

3.2.4. Shale reserves

Fig. 14 shows the most often used names of shale plays on the left, and “fracking” (includes “hydraulic fracturing,” “frac,” and “fracking”) and “shale gas” + “gas shale” on the right. We observe peaking of shale-related terms from 2005 to 2015, which was followed by a decline in recent years. In the past 20 years, “Barnett” shale was mentioned more frequently than other shale deposits. In 2019 “Marcellus,” “Eagle” (Ford), and “Barnett” show the same occurrence, about one time per hundred pages. However, the term “fracturing” does not show such a fast decline. Despite the fact that the names of gas shale deposits reduced in use in the past three years, words that relate to the development and description of these deposits (“fracking,” “TOC” - total organic carbon, “unconventional”) do not show a decrease in use.

It is curious that in 2018, we observe an increase in the mention of the words “student,” “faculty,” and “researcher," Fig. 15, left. Does this mean that the number of academic impacts has grown in the last year? You may notice the peaking of the word “engineer” after peaking of the word “student.” We

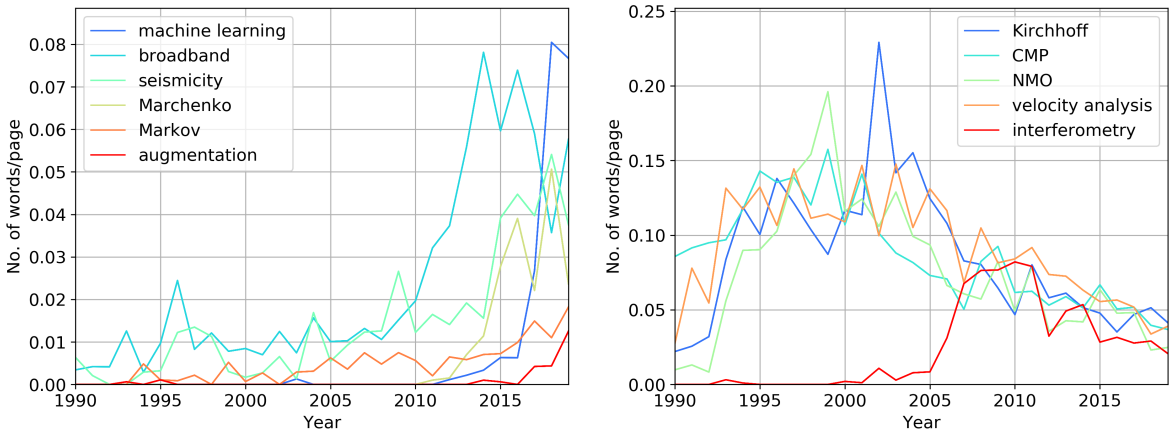

Figure 13. Trends in seismic data processing, terms, and algorithms that start to grow in usage (left) and decline in occurrence (right).

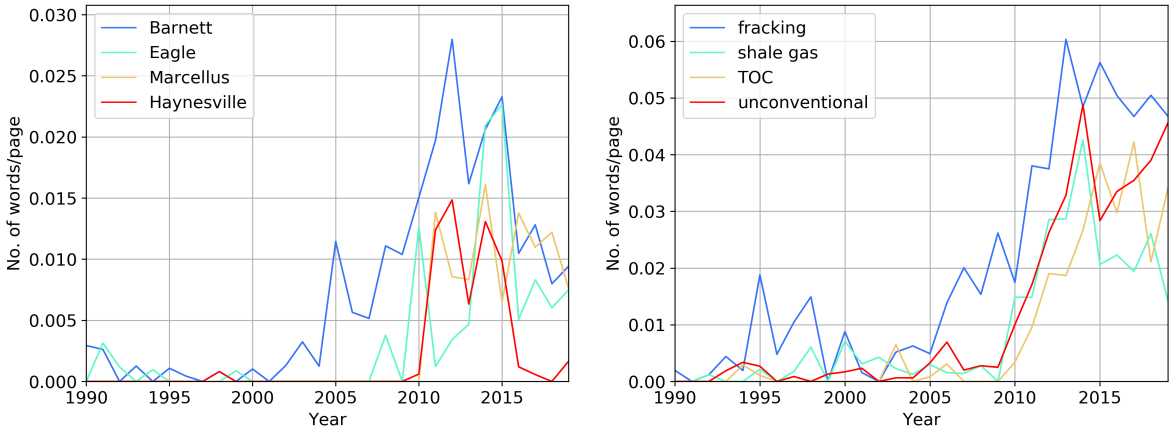

Figure 14. Most frequently mentioned shale reserves; change in usage of hydraulic fracturing and shale gas.

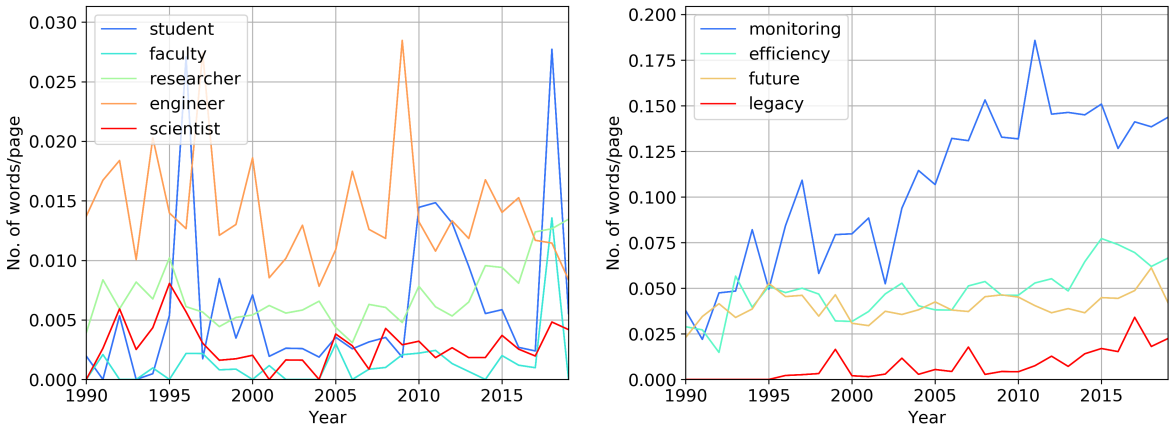

Figure 15. Change in words usage over time.

observe the growth in frequency of the word "researcher" word in the past ten years, and it appeared more often then engineer in 2019. During the 1990s, we see more of the word "engineer" in comparison to "researcher" and "scientist," in the past decade, the situation has changed, bringing "researcher" to the first place. On the right side of Fig. 15, we observe an increase in the usage of the word "monitoring." For example, this applies to microseismic monitoring and monitoring of the reservoir.

The increased use of the term “monitoring” and “efficiency” indirectly indicates the concentration of researchers on the development of already explored deposits. The term “legacy” primarily refers to old data that is reprocessed using modern methods, including CNN. We used the word “future” regularly in the past 30 years, perhaps, we can agree, the past is over.

3.2.5. Neural networks

We see that usually, the growth in the use of terms is saw-like; it is non-monotonic with individual peaks. Each peak represents the next phase of implementation, new research objects, and new teams that have mastered the method. “Neural networks” show a qualitatively different picture. From 1990 to the beginning of 2000, attempts were made to use neural networks in geophysics, but they were suspended until 2016, in which a rapid growth in the use of this and related terms began. On average, we find a “neural network” phrase on every fourth page of the conference materials. If we observe an increased interest in this topic, then the researchers sincerely believe that using the methods of machine learning can solve many problems of geophysics. Given this context, we pose the question: Is the automation of geophysical data processing the main problem of modern geophysics? The authors believe that the main problem of geophysics is the lack of new research objects, such as hydrocarbon and other mineral deposits. Lack of survey objects is the reason for the increased interest in the development of methods for automatic processing of geophysical data. At the same time, the use of words “monitoring” and “efficiency” is growing, which indicates an understanding of the need for complete extraction of hydrocarbons and the monitoring of developed fields. Fig. 16 shows the appearance of “neural network,” “deep learning,” “artificial intelligence” and “field data,” we use the last phrase for reference as it is always often used. In 2019, “neural network,” occurred more often than “field data.” It had already happened in 1993 and from 1999 to 2001, after that, it declined for a while, but now “neural network,” “deep learning,” and “artificial intelligence” have started to grow again ("artificial intelligence” appeared during the 1980s). The question is: Will the growth continue, or will it decline again like it did in 1993 -1995? The decline in interest in neural networks in early 2000 can be explained by an insufficient amount of computing power to realize the capabilities of the method. Now, technological progress allows us to use neural network methods successfully for face recognition. We also see attempts to introduce them to other areas of life. It is not necessary to be a rocket scientist to understand the reasons for the increasing interest in neural networks in geophysics. Experts want to automate geophysical data processing as much as possible. It remains only to understand whether we need to automate seismic data processing deeply. With time, we will have fewer oilfields to be explored, providing space for monitoring and increasing production efficiency.

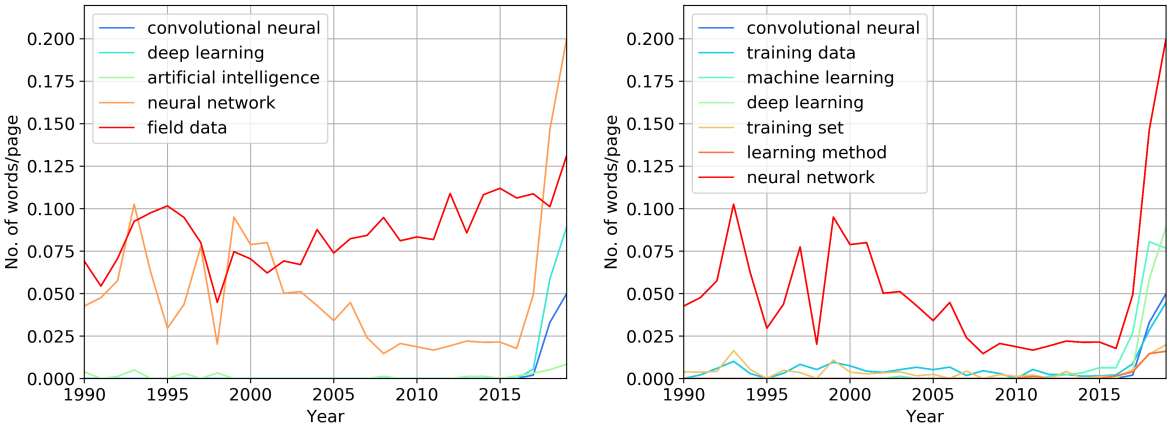

Figure 16. “Neural network” related two-word phrases. We display the phrase “field data” for the reference.

4. Discussion

The emergence of new techniques in geophysics inevitably leads to an increase in the use of terms from this field. The frequency of the occurrence of words can be used to track trends in the equipment, processing methods, math algorithms, and types of resources, including oilfields and the kind of rocks under study. The amount of hidden information is astounding. Such a study is unique because we have at our disposal a history of the development of geophysics. Moreover, it allows us to track exactly how the professional language changes over time.

It is interesting to know the terms that are gaining popularity now and discover the current trends in geophysics. Fig. 17 shows words with the highest growth in occurrence on the left and highest rate of decline on the right. As one can observe, the majority of words that have grown in occurrence relate to the neural network method. Is it possible to assume that these words will continue to gain popularity in the years ahead, and that the topic will remain relevant? For example, the phrases "streamer em" and "receiver deghosting" grew in occurrence at a very fast rate during 2011 – 2015, but since 2015, they decline as quickly as they were growing before. The word "fiber" and "fibre"[4] is increasing in use almost as rapidly; this refers to fiber optics because seismic sensors based on fiber optics are now growing in use, showing their effectiveness in detecting faults filled with geothermal fluids [23], microseismic monitoring during hydraulic fracturing [24] and other applications. The term "distributed acoustic sensing" (DAS) shows good correspondence with the word "fiber" as a DAS is based on fiber-optics, and these terms are closely associated. Here, the use of the word is directly related to the production of the corresponding equipment. For "neural network," one can use the existing computing power. Per contra, the development of optical fiber requires production. However, in 2019, we observe a decline in the usage of the word "fiber." "Wasserstein" (metrics) and (data) "augmentation" have also grown in occurrence in the past three years but not that fast as "Marchenko." Let us conclude, that the lack of research objects forces professionals to develop processing methods and, for example, reprocess legacy data. The picture on the right-hand side in Fig. 17 shows terms that decreased in occurrence in the past four years.

Interestingly, there has been a reduction in the use of the graphics processors by researchers as opposed to seven to eight years ago when the phrase was trending. "Barnett" shale is one of the most well studied, and the authors believe that the fading of interest in it is a natural phenomenon. Curiously, there was increased interest in basalt at the turn of the century, and we observe increased interest in the early 2010s.

Besides "neural network" related terms (Fig. 18) on the left side, we observe an increase in usage of "tight sandstone" and "igneous rock." It is interesting that for 30 years, "igneous rocks" were rarely discussed, except in 2009. In 2018 and 2019, we observe several papers discussing igneous rocks found on Chinese and Brazilian oil fields. Their acoustic and elastic properties must be considered in reservoir characterization [25]. On the right of Fig. 18 one can see two-word phrases that show a decrease in the frequency of occurrence in the past four years. When new research topics appear, new ones will partially or entirely replace old ones since the number of articles is limited every year.

Hill first described Gaussian beam migration in 1990. It is the seismic method that can image steeply dipping reflectors (more than 90 degrees) and will not produce unwanted reflections from the structure in the velocity model [27]. In 1993 at the SEG Annual Conference, we observe several papers reporting usage of beam migration in seismic data processing. In 2001 we notice an increase in the number of occurrence of "beam migration," with the increase in computing capabilities, it became possible to use this method for 3D AVO analysis (Amplitude variation with offset) of small and medium-size 3D seismic surveys [28]. Interest in this method rises two more times, in 2008 and 2015. Frequency peaks appear with enviable regularity every seven or eight years. Moreover, each subsequent peak is higher than the previous one. In 1990, a new method appeared; in 1993, we

[4] "Chiefly British spelling of fiber" [26]

observe testing on synthetic data; in 2001, professionals report the results of processing small and medium volumes of data, in 2007 and 2008, the results of use on large objects in the Gulf of Mexico [29], CGGVeritas. For 25 years, we have seen the emergence of new technologies, testing, and application in field exploration. However, since 2015, we see a decrease in the rate of use of this term. Fig. 18 shows a reduction in the use of other seismic terms and Barnett shale.

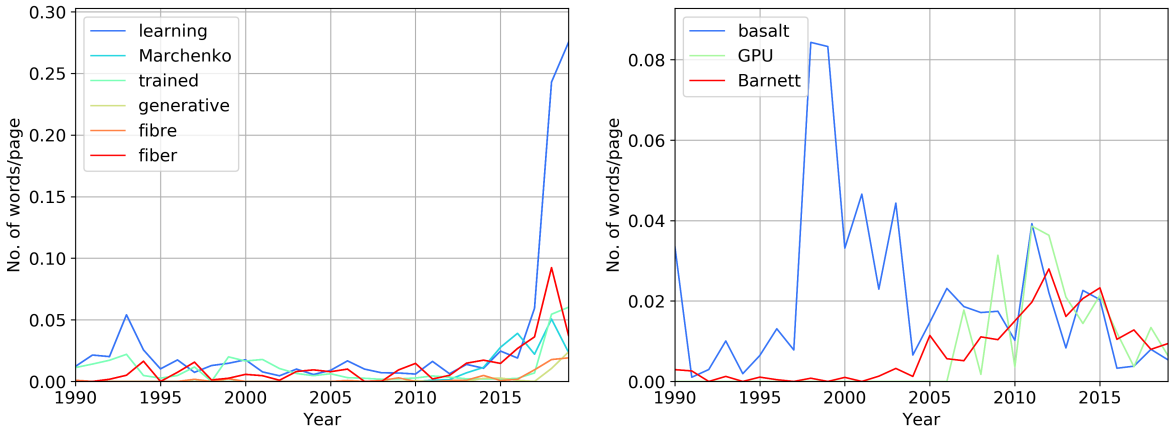

Figure 17. Words that show the highest rate of growth in occurrence (left) and decline (right) in the past four years.

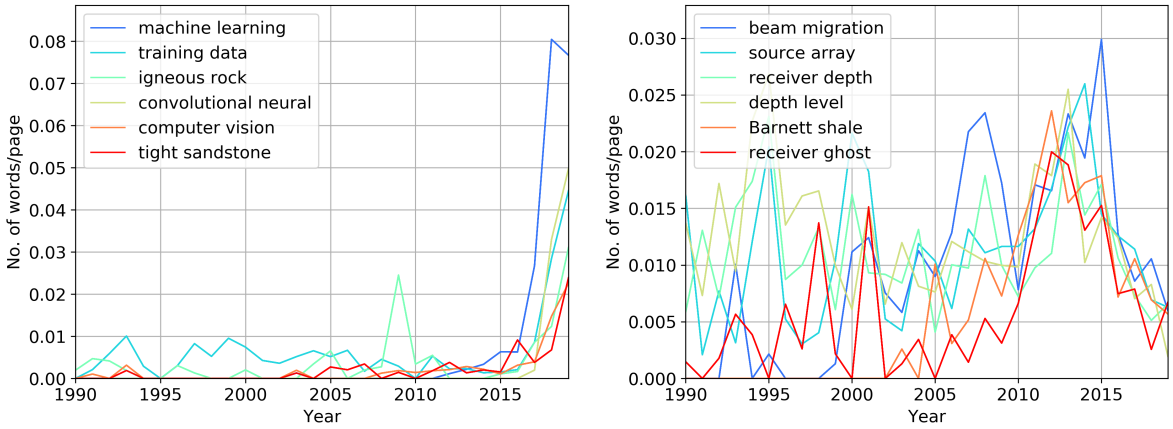

Figure 18. Two-word phrases that show the highest rate of growth in occurrence (left) and decline (right) in the past four years.

Let us consider the most growing and declining three-word phrases, Fig. 19. "Convolutional neural network" (CNN) shows the fastest growth; the second one is "distributed acoustic sensing" (DAS), which is related to the fiber-optic measurement system. In the recent few years, researchers are using CNN to perform "seismic facies classification," which is why we observe an increase in usage. We also see a relative increase for "ground penetration radar," however, we see this term more often during the 1990s and early 2000. The right graph of Fig. 19 shows a decrease in the use of specific seismic terms as for the case of two-word phrases and the names of the shale deposits. From 2010 to 2019, we observe an increase and decrease in interest in the phrase "towed streamer EM." Towed streamer electromagnetic systems allow one to collect data at a high rate and over huge survey areas [30]. It is necessary to have significant objects to survey broad areas. Nowadays, there are less large-scale oil exploration projects, so the researchers use the corresponding terms less often.

It would be interesting to trace how the different methods are developing in geophysics, electrical exploration methods, petrophysics, engineering geophysics. For this reason, it is worthwhile to study

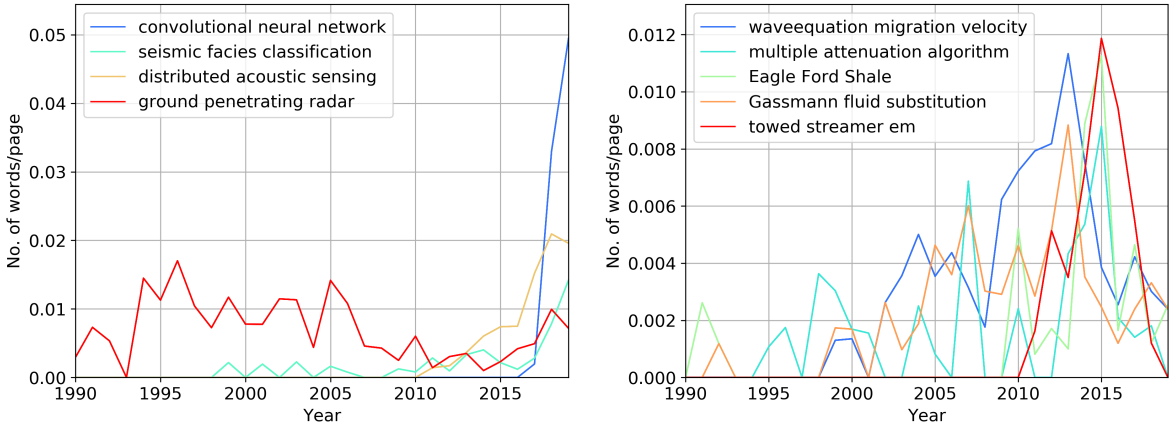

Figure 19. Three-word phrases that show the highest rate of growth in occurrence (left) and decline (right) in the past four years.

the materials of conferences and publications of other journals with a different specialization. Research of conference materials of other societies (SPWLA, EAGE, SPE) will provide a complete picture of the oil industry development.

The authors believe that the resulting database of industrial organizations and universities can be used to study the oilfield services and oil production market, or by students when searching for a place to study. We encourage readers to use our data available online [5]. The data includes the filtered word lists with the frequency of use each year, the number of pages, and the average number of co-authors. Thus, the reader will be able to conduct their research, test their hypotheses or assumptions.

5. Conclusions

We analyzed 24,500 papers, including 127,900 pages consisting of 57 million words, or more than 383 million symbols. Academic institutions from 86 countries and more than 2400 industrial companies contributed to the SEG Annual Conference from 1982 to 2019. The USA academia has the most significant impact in the proceedings of the SEG Annual Conference during the whole observation period. We observe that the number of papers from the Chinese academy is growing, and it is almost equal to those of the USA. The activity of the companies at the SEG Annual Conference shows their economic condition, annual reports by CGG and ExxonMobil and other companies confirm this statement. Depending on the financial situation on the market, and the price of oil, the contribution of the academia and industry by publications changes in time. In 2018 we observed more abstracts from the academia, but in 2019 the number of publications from academia and industry were very close. In 2019 the most significant number of publication in the industry was made by BGP and PetroChina. The average number of authors per paper continues to grow over time in agreement with the global trend of Earth and Planetary science, but at a slower rate.

Alteration in the professional language reflects the change in the industry and science. Over the 30 years, the objects and geophysical methods changed slightly. There has been an increased interest in “shales” in the last ten years. In the past six years, the frequency of the use of the word “shale” has been falling, but the use of the phrases “unconventional,” “TOS,” “hydraulic fracturing” has not decreased in recent years. At the same time, new methods of processing and capturing data appeared, and this led to a change in language. "Neural network” and related disciplines show the fastest growth in the last two years. The authors doubt that growth will continue at the same rate as the term “neural network” is already used more than “field data.” More likely, “neural network” related topics will occupy its niche in geophysics for the coming years. We see an increase in the use of the words “Marchenko,” “seismicity,” and “broadband.” We also observe the rapid growth of the word “fiber,” which is more likely related to fiber optic sensing systems. Supposedly, we will see more

projects on "monitoring" of oil and gas fields and increasing production "efficiency," while there will be less work on the exploration of new oil and gas fields.

Author Contributions: Data mining and processing, software development, original draft preparation - Timofey Eltsov; software development and analysis, review and editing of the draft - Maxim Yutkin; supervision, project administration, historical analysis, review, and editing of the paper - Tadeusz W. Patzek.

Funding: Dr. Eltsov was supported by the KAUST Magnetic Sensor project, REP-2708.

Acknowledgments: Authors appreciate the responsiveness of the SEG team for permission to use digital data and especially SEG Digital Publications Manager, Jeno Mavzer, for the useful advice and help. The authors are grateful to their colleagues, and especially to Dr. Thomas Finkbeiner, for valuable and vital research recommendations. The authors thank Dr. Sergey Yaskevich for consultations on exploration seismic. The authors are grateful to Ilya Kolganov for the useful advice on the design of the graphs. We also would like to acknowledge Dr. Charles Russell Severance for an informative Python course.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASCII	American standard code for information interchange
AVO	Amplitude Variation with Offset
BP	BP plc., formerly The British Petroleum Company and BP Amoco
CGG	Compagnie Générale de Géophysique
CNN	Convolutional Neural Network
CMP	Common Mid Point
CSEM	The Controlled Source Electromagnetic
DAS	Distributed Acoustic Sensing
EAGE	European Association of Geoscientists and Engineers
EM	Electromagnetic
FWI	Full Waveform Inversion
GDP	Gross Domestic Product
GERD	Gross domestic Expenditure on Research and Development
GPU	Graphics Processing Unit
HTML	HyperText Markup Language
NLTK	Natural Language Toolkit
NMO	Normal Moveout
PDF	Portable Document Format
PIL	Python Imaging Library
PSDM	Prestack Depth Migration
RTM	Reverse Time Migration
R&D	Research and Development
SEG	Society of Exploration Geophysicists
SPE	Society of Petroleum Engineers
SPWLA	Society of Petrophysicists and Well Log Analysts
TXT	Text file
TOC	Total Organic Carbon
USA	The United States of America

Appendix A. The total number of publication by country

#	Country name	The total number of publications
1	United States of America	5154.07
2	China	1649.95
3	Canada	952.47
4	Netherlands	608.58
5	France	546.53
6	United Kingdom of Great Britain and Northern Ireland	462.14
7	Germany	330.39
8	Australia	302.54
9	Brazil	264.0
10	Japan	205.62
11	Russian Federation	178.75
12	Norway	176.36
13	Italy	164.11
14	Saudi Arabia	145.46
15	Korea, Republic of	138.62
16	Switzerland	113.01
17	India	99.08
18	Mexico	62.15
19	Denmark	57.58
20	Israel	55.69
21	Taiwan	49.44
22	Sweden	41.45
23	Venezuela (Bolivarian Republic of)	28.01
24	Argentina	25.78
25	South Africa	24.53
26	Nigeria	23.96
27	United Arab Emirates	19.75
28	Czechia	19.34
29	Spain	17.86
30	Egypt	16.52
31	Singapore	14.98
32	New Zealand	14.77
33	Romania	13.7
34	Indonesia	13.42
35	Malaysia	12.66
36	Greece	12.06
37	Portugal	11.74
38	Ukraine	11.08
39	Colombia	10.25
40	Finland	9.8
41	Austria	6.92
42	Iran	6.78
43	Belgium	5.43
44	Slovakia	5.35
45	Ireland	5.26
46	Poland	5.22
47	Turkey	5.2
48	Serbia	4.5
49	Thailand	4.34
50	Jamaica	4.32

References

1. Glauner, P.; Valtchev, P.; State, R. Impact of Biases in Big Data. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 25-27 April 2018; pp 645–654.
2. Kaplan, R.; Chambers, D.A.; Glasgow, R.E. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *CTS Journal* **2014**, *7*, 4, 342–346. doi:10.1111/cts.12178.
3. SEG Technical Program Expanded Abstracts. Available online: https://library.seg.org/series/segeab (Accessed on 2 December 2019).
4. OnePetro online library. Available online: https://www.onepetro.org (Accessed on 2 December 2019).
5. Eltsov, T. Data for SEG Annual Conferences analysis, 1982 - 2019. Available online: https://github.com/ANPERC-source/SEG_Annual (Accessed on 2 February 2020).
6. American Higher Education Hits a Dangerous Milestone. Available online: https://www.theatlantic.com/politics/archive/2018/05/american-higher-education-hits-a-dangerous-milestone/559457/ (Accessed on 2 December 2019)
7. Paper authorship goes hyper. Available online: https://www.natureindex.com/news-blog/paper-authorship-goes-hyper (Accessed on 17 October 2019).
8. UNESCO. *UNESCO SCIENCE REPORT, Towards 2030*; Report; United Nations Educational, Paris, France, 2015.
9. Ni, X. China's research & development spend. *Nature* **2015**, *520*, S8-S9. doi:10.1038/520S8a.
10. Global Economic Data, Indicators, Charts& Forecasts, CEIC. Available online: https://www.ceicdata.com (Accessed on 10 February 2020).
11. Compagnie Générale de Géophysique *ANNUAL REPORT 2015*; Report; CGG: Chicago, Ill., USA 2015.
12. Compagnie Générale de Géophysique CGG completes its exit from marine acquisition. Available online: https://www.cgg.com/en/Investors/Press-Releases/2020/01/CGG-Completes-its-Exit-from-Marine-Acquisition (Accessed on 2 February 2020).
13. Schlumberger. *2014 Annual Report*; Report; Schlumberger: Paris, France 2015.
14. Saudi Aramco. Saudi Arabian Oil Company, Consolidated financial statements for the year ended december 31, 2018. Available online: https://www.saudiaramco.com/-/media/publications/corporate-reports/saudi-aramco-results-2017-2018-full-financials.pdf (Accessed on 11 February 2020).
15. ExxonMobil. *Summary Annual Report 2005*; Report; ExxonMobil: Irving, TX, USA, 2005.
16. ExxonMobil. *Summary Annual Report 2014*; Report; ExxonMobil: Irving, TX, USA, 2014.
17. Kroode, F.; Bergler, S.; Corsten, C.; Maag, J.W.D.; Strijbos, F.; Tijhof, H. Broadband seismic data — The importance of low frequencies. *Geophysics* **2013**, *78*, 2, WA3–WA14. doi:10.1190/GEO2012-0294.1.
18. Lomas, A.; Curtis, A. An introduction to Marchenko methods for imaging. *Geophysics* **2019**, *84*, 2, 35–45. doi:10.1190/geo2018-0068.1.
19. Thorbecke, J.; Slob, E.; Brackenhoff, J.; Neut, J.V.D.; Wapenaar, K. Implementation of the Marchenko method. *Geophysics* **2017**, *82*, 6, WB29–WB45. doi:10.1190/geo2017-0108.1
20. Iervolino, I.; Giorgio, M.; Chioccarelli, E. Markovian modeling of seismic damage accumulation. *Earthquake Engineering & Structural Dynamics* **2016**, *45*, November 2015, 441–461. doi:10.1002/eqe.
21. Weir, R.; Lines, L.; Lawton, D.; Eyre, T. The Duvernay Formation : the application of structure and simultaneous inversion for reservoir characterization and induced seismicity. In Proceedings of the SEG Annual Conference and Exhibition, Anaheim, USA, 14-19 October 2018; pp 2372–2376. doi:10.1190/segam2018-2980345.1.
22. Barthwal, H.; Baan, M.V.D. Causative mechanism of microseismicity recorded in an underground mine. In Proceedings of the SEG Annual Conference and Exhibition, Anaheim, USA, 14-19 October 2018; pp 2962–2966. doi:10.1190/segam2018-2980345.1.
23. Trainor-Guitton, W.; Jreij, S.; Guitton, A.; Simmons, J. Fault classification from 3D imaging of vertical DAS profile. In Proceedings of the SEG Annual Conference and Exhibition, Anaheim, USA, 14-17 October 2018; pp 4664–4668.

24. Chakraborty, G.; Chakraborty, D. Detecting microseismic events in downhole distributed acoustic sensing data using convolutional neural networks. In Proceedings of the SEG Annual Conference and Exhibition, San Antonio, USA, 15-20 September 2019; pp 4864–4868.
25. Penna, R.; Araújo, S.; Geisslinger, A.; Sansonowski R.; Oliveira, L.; Rosseto J.; Matos, M. Carbonate and igneous rock characterization through reprocessing, FWI imaging, and elastic inversion of a legacy seismic data set in Brazilian presalt province. *The Leading Edge* **2019**, *38*, 1, 11–19. doi:10.1190/tle38010011.1.
26. Merriam-Webster online dictionary. Available online: https://www.merriam-webster.com/ (Accessed on 16 February 2020).
27. Hill, N.R.; Gaussian beam migration. *Geophysics* **1990**, *55*, 11, 1416–1428. doi:10.1190/1.1442788
28. Huang, S.; Sherrill, F.; Sengupta, M.K. Merits of amplitude preserving Kirchhoff beam migration method for 3D AVO analysis. In Proceedings of the SEG Annual Conference and Exhibition, San Antonio, USA, 9-14 September 2001; pp 1–4.
29. Ting, C.O.; Wang, D. Controlled beam migration applications in Gulf of Mexico. In Proceedings of the SEG Annual Conference and Exhibition, Las Vegas, USA, 9-14 November 2008; pp 368–372.
30. Zhdanov, M.S.; Endo, M.; Sunwall, D.; Mattsson, J. Advanced 3D imaging of complex geoelectrical structures using towed streamer EM data. In Proceedings of the SEG Annual Conference and Exhibition, New Orleans, USA, 18-23 October 2015; pp 904–908.