SPE-190812-MS

Status of Data-Driven Methods and their Applications in Oil and Gas Industry

Karthik Balaji and Minou Rabiei, University of North Dakota; Vural Suicmez, QRI Analytics; Celal Hakan Canbaz, Schlumberger; Zinyat Agharzeyva, Texas A & M University; Suleyman Tek, University of the Incarnate Word; Ummugul Bulut, Texas A&M University-San Antonio; Cenk Temizel, Aera Energy LLC

This paper was prepared for presentation at the SPE Europec featured at 80th EAGE Conference and Exhibition held in Copenhagen, Denmark, 11-14 June 2018.

Abstract

Data-driven methods serve as robust tools to turn data into knowledge. Historical data generally has not been used in an effective way in analyzing processes due to lack of a well-organized data, where there is a huge potential of turning terabytes of data into knowledge. With the advances and implementation of data-driven methods data-driven models have become more widely-used in analysis, predictive modeling, control and optimization of several processes. Yet, the industry overall is still skeptical on the use of data-driven methods, since they are data-based solutions rather than traditional physics-based approaches; even though physics and geology are often part of this methodology. This study comprehensively evaluates the status of data-driven methods in oil and gas industry along with the recent advances and applications.

This study outlines the development of these methods from the fundamentals, theory and applications perspective along with their historical acceptance and use in the industry. Major challenges in the process of implementation of these methods are reviewed for different areas of applications. The major advantages and drawbacks of data-driven methods over the traditional methods are discussed. Limitations and areas of opportunities are outlined. Recent advancements along with the latest applications, the associated results and value of the methods are provided.

It is observed that the successful use of data-driven methods requires strong understanding of petroleum engineering processes and the physics-based conventional methods together with a good grasp of traditional statistics, data mining, artificial intelligence and machine learning. Data-driven methods start with a data-based approach to identify the issues and their solutions. Even though data-driven methods provide great solutions on some challenging and complex processes, that are new and/or hard to define with existing conventional methods, there is still skepticism in the industry on the use of these methods. This is strongly tied to the delicacy and sensitive nature of the processes and on the usage of the data. Organization and refinement of the data turn out to be important components of an efficient data-driven process.

Data-driven methods offer great advantages in the industry over that of conventional methods under certain conditions. However, the image of these methods for most of the industry professionals is still fuzzy. This study serves to bridge the gap between successful implementation and more widely use and acceptance

of data-driven methods, and the fuzziness and reservations on the understanding of these methods in the industry. Significant components of these methods along with clarification of definitions, theory, application and concerns are also outlined in this study.

Introduction to Data Driven Methods

Data-driven methods, in general, provide a set of tools and techniques to integrate various types of data, quantify uncertainties, identify hidden patterns, and extract useful information. This information is used to predict future trends, foresee behaviors, and answer questions, which are not possible by conventional models. Models in the oil and gas industry can be classified into three main categories as mathematical, physical and empirical. Mathematical models represent an approximation to the physical world and are based on the first principles such as mass and energy, conservation of momentum, etc., or on empiricism or a combination of both. Although, the combination of first principle models with the constitutive equations are enough to generate the model structure in a range of operating conditions, they are incapable of providing good enough solutions in certain conditions (Saputelli et al., 2003). On the other hand, empirical models, although easy to develop, cannot be generalized and may not be accurate. Conventional modelling approaches in the industry, are usually lengthy procedures based on trial and error, which involve iterative refining until the desired results are obtained. These models cannot handle very complex relations and generally require a number of assumptions and simplifications. They are also unable to account for noise and missing data.

Technological advances in the oil and gas industry and increasing internet usage have made it possible to generate staggering amount of data, generally called Big Data, which is often not efficiently and effectively used. Big Data is characterized by 3V's, namely, volume, variety, and velocity (Mishraand Datta-Gupta, 2017). Big Data is becoming an essential part of the oil and gas industry (Holdaway, 2014; Saputelli, 2016). The increasing complexity of oil and gas systems due to non-linearity and high level of uncertainties and the large volumes of data generated, call for more sophisticated models that can turn raw data into actionable knowledge and are able to represent the complex relationships among the system state (input, internal and output) variables.

Data mining, a major component of data driven methods, can offer great benefits to the oil and gas industry by extracting hidden predictive information from the large and/or complex databases. Data mining integrates machine learning and pattern recognition algorithms with statistical and visualization techniques to convert data into meaningful and comprehensible knowledge. These techniques use past and present information to discover previously undetected patterns in the data to construct predictive or descriptive models with good generalization abilities. Data-driven modelling (DDM), provides the methodology for analyzing and discovering the relationships among the system state variables without the explicit knowledge of the physical behavior of the system. DDMs enable proactive, knowledge-driven decision making towards preventing issues and solving complex problems during a reservoir life cycle, and facilitates identification of best practices in the industry. Data-driven modeling focuses on two empirical models. Computational Intelligence which involves artificial neural networks, fuzzy rules-based systems, Genetic Algorithms, and Machine Learning models based on the theoretical foundations used by Computational Intelligence. These two methods mainly help to construct the data-driven models in Oil and Gas Industry (Holdaway, 2014). In a typical Data Analysis Cycle, data collection and management is the first step. In this step, data is collected from various resources and in different forms and it is prepared for analysis. Next, the variables required for the analysis are decided through exploratory data analysis (EDA) and their importance and the connections among them are explored.

EDA techniques have been adopted into data mining processes designed for analyzing large amounts of data. Data mining in its most fundamental form, is defined as extracting interesting, nontrivial, implicit, previously unknown and potentially useful information from data. During a data mining process predictive

or descriptive models are built using machine learning techniques. Descriptive mining tasks describe the general properties of the data and find human-interpretable patterns that characterize the data. Predictive mining tasks perform inference on the current variables to predict unknown or future values of other variables. Machine learning algorithms work by taking a set of input values in the form of a vector written as X, with elements x_i, where i runs from 1 to the number of input dimensions, n, and producing an output (response) for that input vector. The output vector Y, contains elements y_j, where j runs from 1 to the number of output dimensions, k. Machine learning algorithms can be classified into 4 major categories, as shown in Fig. 1.

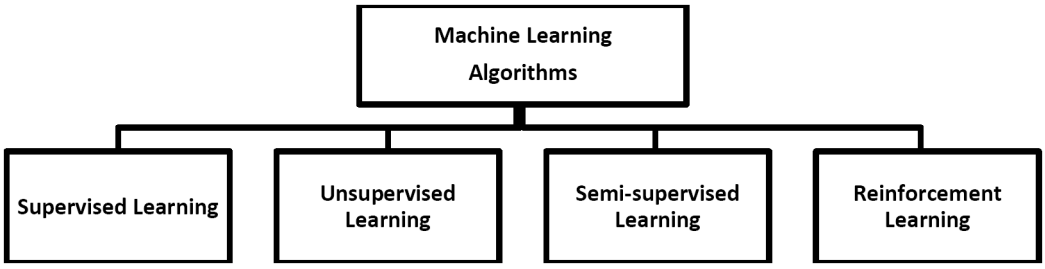

Figure 1—Different Categories of Machine Learning Techniques

In supervised learning, a training set of instances (examples), where each instance contains a number of predictor attributes (x_i) and an associated response (y_j), representing the class (label) of the instance (classification). Using this training set, the algorithm learns the patterns in the data and generalizes to correctly classify the new inputs. An unsupervised algorithm is not provided with a response attribute. Instead the algorithm identifies the similarities between the inputs and categorizes the instances based on these similarities (clustering). A semi-supervised learning algorithm is trained by a training set that contains both labeled and unlabeled data. The semi-supervised algorithm can in some cases achieve better performance than supervised learning without requiring the data to be fully labeled. In reinforcement learning, the algorithm receives feedback on where it went wrong but the instruction on how to correct it is not provided; so the algorithm trains itself continually using trial and error.

Data-Driven Modelling Techniques

Some of the most commonly used data-driven modelling techniques and a brief introduction to each of these algorithms are presented in this section.

Linear Regression

One of the first and important modeling techniques is linear regression which can model the relationship between a dependent variable and one or more explanatory (predictor) variables. In a dataset containing only one predictor variable (X_1, Y_1), (X_2, Y_2),…(X_n, Y_n), where Y depends on X, a linear model that fits to the data is given in the following form:

$$Y = a + bX + \varepsilon\# \tag{1}$$

Here a and b are regression parameters that need to be calculated using the given data, and ε is the error.

The best fitted line to the data is the one for which the total prediction error or the sum of the squared prediction errors between the given data and the predicted values is as small as possible (Haan, 1986).

$$\min r(\hat{a},\hat{b}) \ = \sum_{i=1}^{n} \left(Y_i - \ Y_p(X_i)\right)^2 \ = \sum_{i=1}^{n} \left(Y_i - \hat{a} - \hat{b}X_i\right)^2 \# \tag{2}$$

Where, Y_p is the predicted value.

If there are more than one predictor variables, the regression process is called a multiple linear regression and it can be modelled as:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i \,, \# \tag{3}$$

Principal Component Analysis

Most machine learning and data mining techniques may not be effective for high-dimensional data. Therefore, some preprocessing is required to reduce the dimensions of data and deal with what is commonly known as "curse of dimensionality". Dimensionality reduction techniques work by identifying an optimal subset of attributes or features according to an objective function or by reducing the number of features by creating linear combinations of the original attributes. Principal component analysis reduces the dimensions of the data by performing a linear transformation on the data by rotating the feature space, so the data lies along the directions of maximum variation. In this algorithm, a matrix X of the input data vectors is formed which comprises *n* observations of *m* dimensions. The covariance matrix can be calculated using the transpose of the X matrix as:

$$\mathrm{C_x} = \frac{(\mathrm{X} - \overline{\mathrm{X}})\,(\mathrm{X} - \overline{\mathrm{X}})^{\mathrm{T}}}{(\mathrm{n} - 1)} \tag{4}$$

The covariance of the transformed data C_y can then be calculated by finding the matrix P of the first k eigen vectors of matrix C_x such that:

$$\mathrm{C_y} = \mathrm{P C_x P^T} \tag{5}$$

Decision Trees

Decision tree is a commonly used data analytics tool which is widely used in software industry, pharmaceutical industry, law and business. Decision trees are also one of the most visually effective and understandable data mining techniques. Decision Tree algorithm is a supervised classification and prediction technique, which relies on an "attribute" or a group of "attributes" to create models of the data (Fig. 2).

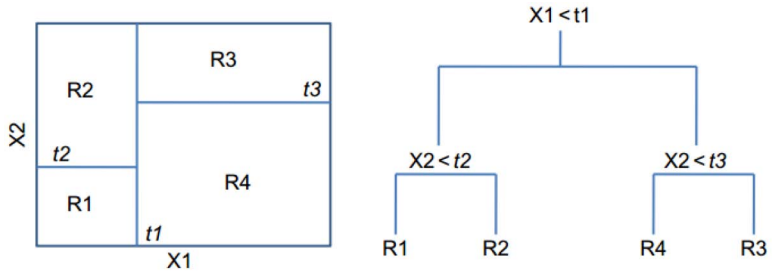

Figure 2—Tree-based modeling (Mishra and Datta-Gupta, 2017)

The initial dataset, is fed into the algorithm in the first node known as the root node. The algorithm analyzes the data points and checks the best attribute to split the data in order to classify data accurately into their respective classes. The most commonly used splitting criteria for choosing the best partitioning attribute are Gini Index, Entropy and Classification Error. The result of each split can be two 2 or more "child" nodes and the node which underwent the splitting process is known as the "Parent" node. Most of these methods rely on the amount of information that each child node gains as compared to its parent node. The algorithm recursively checks the point of split by rating the various possible split points to choose the

best one. The splitting can be stopped before the tree grows into its full length by defining a stopping criterion such as reaching class homogeneity or reaching the minimum number of instances for a non-terminal node. The tree can also grow to its full length and then a procedure called pruning is used to trim the decision tree in a bottom-up fashion. Pruning primarily reduces redundancy in the model, makes the model more impactful with each growth level and also removes repetitions.

The goal in decision trees is to continuously partition the data until all or majority of the data points within each node belong to a specific class. However, a major pitfall of decision trees is their inability to perform well when there is limited data in the training set. The output of this algorithm is very susceptible to any changes in the training data and it could be biased depending on the number of data points from each class in the dataset. Random forest technique, a collection of multiple trees grown in parallel from the same dataset, can overcome many limitations of a single decision tree model. Random Forests are pretty good with handling noise, great with large multi-attribute data (scalability is not an issue), highly accurate and good at tuning the algorithm. (Hastie et al., 2008).

Support Vector Machines

Support Vector Machine (SVM) is an important machine learning model that is used for supervised classification of the data. SVM as an algorithm can be viewed as a clustering tool with features of decision trees. The algorithm in SVM is trained in order to choose a hyperplane that best separates the two classes as seen in Fig. 3. When classes are linearly classifiable a plane is chosen which is on the exact center between the two classes, such that $h(x)<0$ for all points below the hyperplane and $h(x)>0$ for all points above the hyperplane. All points above the hyperplane are given a +1 value and all points below are given −1 values, thus increasing the classification efficiency. The hyperplane has a weight value, which defines its orientation and a bias value, which defines its deviation from the origin point. The data points which have the maximum effect on the separating hyperplane are the support vectors. The SVM algorithm performs very well with data with high dimensionality but it is vulnerable to noisy data and it can get very complex with high processing time.

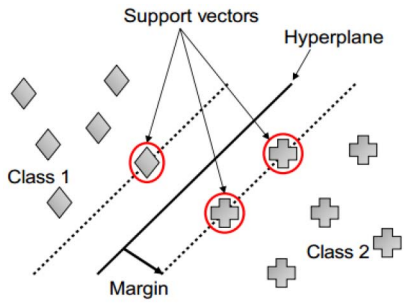

Figure 3—Hyperplane in SVM (Mishra and Datta-Gupta, 2017)

SVM is primarily used for classification but it is also used for regression. Vapnik (1997) used SVM for regression and the method is called support vector regression (SVR). To classify nonlinear boundaries, a nonlinear kernel function that transforms nonlinear hyperplane to linear hyperplane in higher-dimensional space can be used as seen in Fig.4. Kernel function can model the data vectors in a format to make the data fit linearly in a higher dimension.

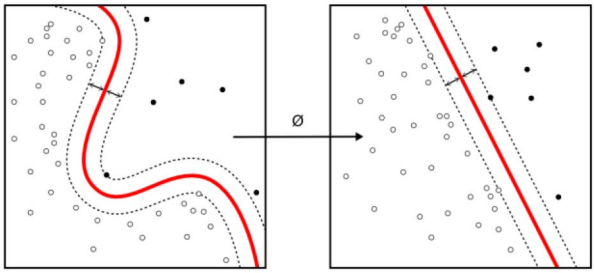

Figure 4—Nonlinear Kernel function (Alisneaky, 2011)

Artificial Neural Network

Artificial Neural Networks can be described as information processing systems that shows similarities with biological neural networks. In other words, artificial neural networks are a rough approximation and simplified simulation of biological neural networks. ANNs are capable of developing transformations, associations and mappings between data. Neural networks are very effective in handling non-linear relations in data and can perform well in extremely complex functions. ANNs have become the go-to solution for a wide range of problems in the oil and gas upstream and downstream. Neural networks are used when the problem is too complicated for mathematical modelling or when there are missing data.

A neural network consists of a number of layers containing input parameters (x_1, x_2,… x_n), a collection of independent neurons in hidden layer(s) and a set of outputs (Fig. 5).

The neurons in these layers are all connected using some weights. Each neuron sums its weighted inputs and applies an activation function to generate the output. The last layer contains a number of outputs, which are compared to the target values (the correct response for a set of input parameters) in order to determine the error. A cost function is defined to quantify this difference computed as the sum-of-squares difference between the network outputs and the target values. The principle of "Gradient Descend" is then used to adjust the weights in order to minimize this cost function.

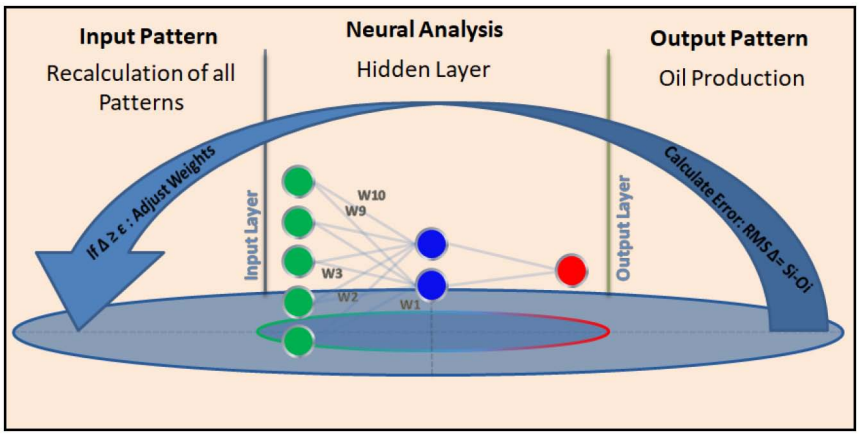

Figure 5—ANN Back-propagation Model (Three Layers) (Holdaway, 2013)

General structure of a back propagated ANN algorithm has the followings steps:

1. Start the algorithm by assigning random weights to the connections between neurons.
2. Obtain the activation value of each neuron using the neuron inputs and their connection weights.
3. Compare the output from the network with the correct target responses to determine the error.
4. Moving backwards, adjust the connection weights based on the error rate, so that the output from the next iteration is closer to the desired response.

When constructing the ANN, the number of hidden layers and nodes, rate of learning, and iteration numbers are the parameters need to be optimized.

Fuzzy rule-based systems (FL)

The idea of "fuzzy rules based systems as well as the term of "fuzzy" was first described by Zadeh in 1965. A fuzzy logic model consists of fuzzy sets, which are formed by the functions of approximate reasoning as well as non-statistical uncertainty. Uncertainties are mainly caused by information insufficiency and it can be described as the product of excursiveness. The duty of a fuzzy logic system is to model the uncertainty that creates the complexity and imprecision. The output product of an event in a random process strongly depends on a chance. Herein, probability theory is a good tool to handle the problem when uncertainty is the product of the randomness of events.

Fuzzy sets may be viewed as an extension and generalization of traditional sets. Fuzzy sets have a membership criterion ranging from 0 to 1. Items in a fuzzy set do not need to completely belong to that specific set but can also belong to other sets in varying membership criteria (partial membership). The concept of vagueness is introduced by using fuzzy sets through elimination of sharp boundaries existing between sets (or classes) (Sivanandam et al., 2011). Fuzzy sets can accomplish operations such as union, intersection and compliment. Similar to traditional relations between sets, there exists the concept of fuzzy relations. Based on this concept, through Cartesian product of two sets, everything is related to some existent or unrelated. The properties of commutativity, associativity, distributivity and other properties hold true for fuzzy relations. The membership function of a fuzzy set defines the fuzziness irrespective of its constituent elements. Membership functions can be assigned to a traditional set through a fuzzification process called "kernel of fuzzification" and then followed by procedures like rank ordering, neural network algorithms or genetic algorithms.

Fuzzy set theory is a popular tool in the O&G industry (Holdaway, 2014). Hydrocarbon reservoirs are complex systems with high percentage of uncertainties due to lack of information. Fuzzy set theory is a magnificent tool that describes the kind of uncertainty associated with vagueness, imprecision and lack of information. In the light of fuzzy variables (low, average, high), approximate reasoning process can be performed, where decisions are made by fuzzy set operators (and, or).

Genetic Algorithms (GA)

Genetic algorithms, first studied by John Holland in 1960s, are basically described as a heuristic search technique that tries to emulate the human evolution to solve an optimization problem. Genetic Algorithm mechanism generally consists of various steps such as initialization, assignment, selection, crossover and mutation (Fig. 6).

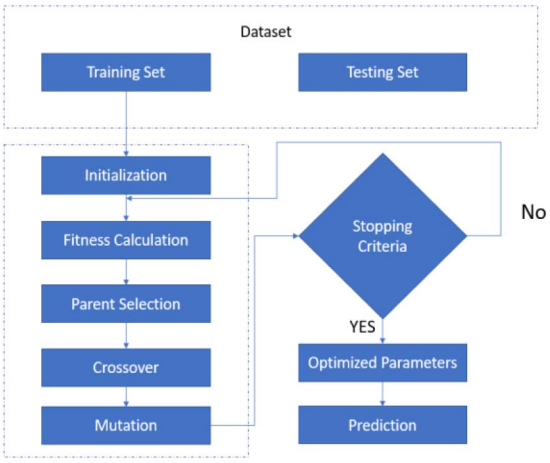

Figure 6—Genetic Algorithm Flowchart (after Bian et al., 2016)

The algorithm starts by randomly generating a population of ***k*** chromosomes, each of which represents a solution. These initial chromosomes create ***k*** offspring through cross over and mutation as shown in Fig. 7.

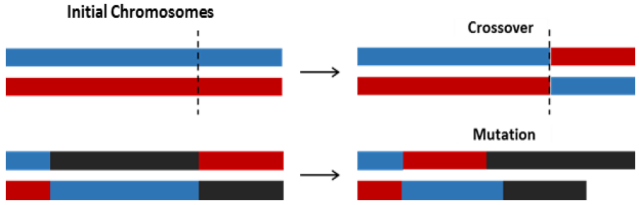

Figure 7—Generating offspring chromosomes by crossover and mutation

In the next step, using a fitness function the best ***k*** chromosomes are detected from the generated offspring. These processes continue iteratively until the best set of chromosomes that satisfy a threshold, are created. GAs are able to deal with complex multidimensional and discontinuous problems. Because GAs do not check every possible combination of genes, they provide a fast and accurate solution for problems with a lot of data.

Bayesian Belief Networks (BBN)

Bayesian Belief Networks is a supervised classification and regression algorithm which has a structure similar looking to decision trees. However, the growth of the network is fueled by best priori probability distribution. Bayesian Belief Networks has its root deeply invested in the principle of conditional probability and cause-effect relationships between multiple random variables. BBN is a probabilistic graphical model (PGM) consisting of a set of nodes and connections, (edges), which represents a decomposition of a large probabilistic domain into weakly connected subsets via conditional independency. The network is usually pre-designed by the user and then the system depending on the instances learns the classification depending on Bayes theory. A Naïve Bayes Classifier, is a simple Bayesian algorithm, which assumes that input attributes are conditionally independent given the class. Unlike Naïve Bayesian theory it is possible to frame interdependency between instances using the parent-child combination as seen in decision trees. While building the probability distribution of classes of a node, it takes into account all the parent nodes of the network too. The nodes represent all the possible values/states of the variables and the edges represent causal relationships between variables. The connection between nodes represent the conditional probability between them, meaning how the state/value of one node can affect the probability distribution of the other node. Each node, is accompanied by a probability distribution table which defines the state of the node. Suppose we have an outcome $B \in \{1, 2, \ldots, b\}$ and an input vector $A = (a_1, a_1, \ldots, a_n)$. Baye's Theorem (Baye's rule) gives the conditional probability as depicted in Eq. 6.

$$P(B|A) = \frac{P(A|B).P(B)}{\mathrm{P}(A)} \tag{6}$$

Where, $P(A|B)$ is the likelihood, $P(B)$ is the prior probability of the class and $P(B|A)$ is the posterior probability of class b. The goal is to find the value of B that maximizes $P(\mathrm{B}|a_1, a_1, \ldots, a_n)$, so the posterior probability for all values of B is calculated using the Bayes Theorem. In other words, the predicted class essentially depends on the likelihood of that class taking its prior probability into account. Fig. 8 displays an example BBN applied in assessment of blowout consequences for Measured Pressure Drilling Operation.

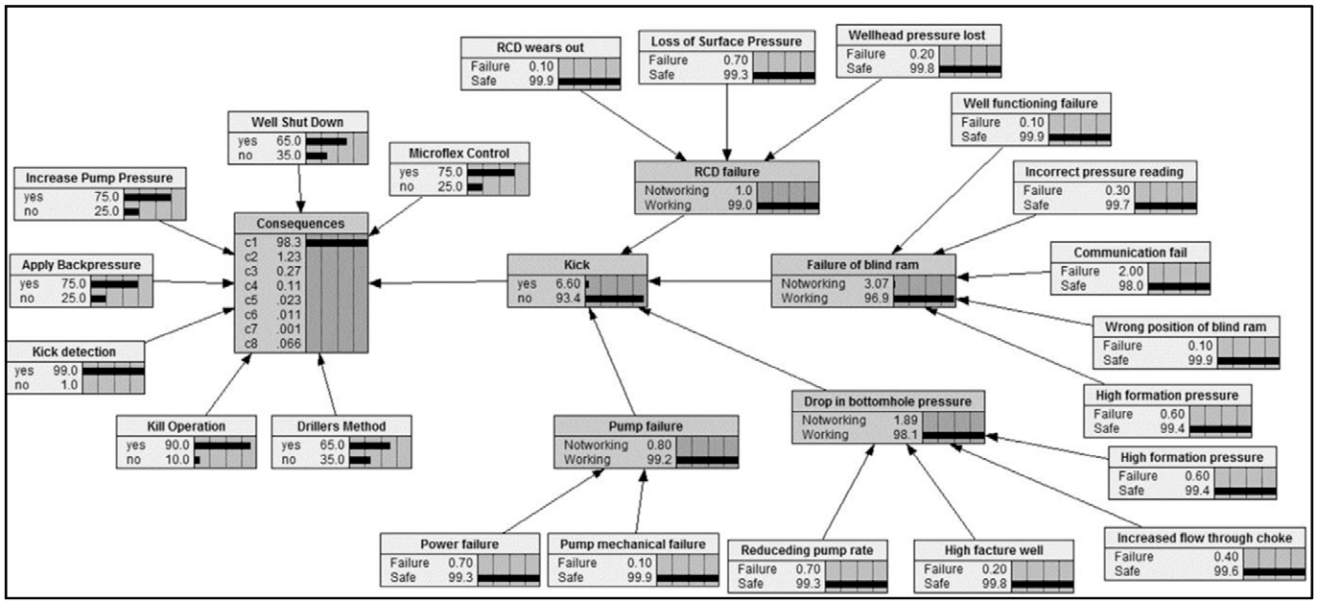

Figure 8—BBN for assessment of blowout consequences for Measured Pressure Drilling Operation (Bhandari et al., 2015)

Data-Driven Methods in Oil & Gas Industry

Mathematical models have been widely used in the oil and gas industry. For instance, Taner et al. (1979) developed a mathematical model that explains the application of complex trace analysis to seismic data and enables to use it in geologic interpretation. However, mathematical models have tough limitations and are harder to simulate. Data-driven methods have also been used by several authors in the oil & gas industry. Application areas are mainly reservoir management and simulation, production and drilling optimization, real-time drilling automation, and facility maintenance (Holdaway, 2014). In this section, some applications of the discussed data driven methods in different disciplines in the industry are presented.

Subsurface Characterization & Petrophysics

The effectiveness of fuzzy logic and neural networks in fractured reservoir characterization was studied by Ouenes in 2000, by applying three different steps to evaluate the performance of two different models. By using this method, Ouenes showed that the effect of each model input on fractures can be described and the parameters that may have a strong correlation with fractures can be identified by using fuzzy curves (Zhang et al., 2018). Al-Anazi et al. (2012) completed a study using support vector regression which is an extension of SVM for appropriate prediction of porosity and permeability values for a field. Behnoud Far, et al. (2017) showed how random forest tree algorithm backed by naïve Bayesian operation can be used for field permeability analysis. Ozkaya (2008) demonstrated the use of decision trees for intersecting and near-well fracture corridors. Gas PVT properties as well as oil-gas interaction have also been modeled using SVM and decision trees as shown by Chamkalani et al. (2013), El-Sebakhy (2009), Tohidi-Hosseini (2016) and Ahmedi (2016). One of the major uses of artificial intelligence includes analyzing logs and generation of missing logs tracts. Cranganu et al. (2013) implemented a Support Vector Regression system in order to generate a sonic log to analyze over-pressured zones at the Anardarko Basin. Akande et al. (2017) developed a support Vector Regression algorithm backed by evolutionary algorithm in order to obtain best hydrocarbon estimates from logs obtained from a reservoir. Estimation of Total Organic Content of a reservoir from log data was another example of the application of machine learning (Tan et al., 2015). Masoudi et al. (2012) developed a supervised BBN algorithm, which learned from logs to generate a model for picking out reservoirs, without user specified cut-offs. Anifowose et al. (2015) proposed an ensemble SVM method in a way similar to random forest tree algorithm for prediction of porosity and permeability values.

Drilling

There has significant progress in the field of drilling especially in the areas of risk control, controlled rate of penetrations and so on. Ahmedi (2016) used SVM to simulate the performance of different types of drilling fluids' rheology under different environmental conditions. Furthermore, cross verification, proved a good match between the prediction and lab results. Fatehi et al. (2017) developed a transducive support vector machine system for mapping potential drilling targets during exploration using deposition information. Zhang et al. (2018) designed a BBN for efficient analysis of risk and uncertainty involved in managed pressure drilling. This analysis considers certain factors to measure uncertainty using extra probability parameters. Al-Yami et al. (2016) used BBN to build a drilling expert system based on various reservoir and fluid parameters. A system to predict the circumstances leading to an offshore blow-out especially during conducted measured pressure drilling and unbalanced drilling and risk analysis, was created by Bhandari et al. (2015). A similar study, performed by Sule et al. (2018), analyzes the reliability of system controls after simulating kick conditions used in measured pressure drilling. Also, Chang et al. (2018) conducted a study on emergency riser disconnection modules using BBN algorithms. Six criteria of disconnect modules were linked between the BBN system and failure tree study. Similarly, Cai et al. (2012) developed a BBN to study the reliability of Blowout Preventer redundancy in deep sea wells. Kormaksson et al., (2015) used principal component analysis to determine economically feasible locations for new wells. Bakshi (2017) used a novel nonlinear regression model to forecast shale oil well performance such as optimized well locations and completion parameters. Temizel et.al. (2016) examined the elements that influence the performance of vertical and horizontal wells in tight reservoirs.

Production

Production is an area of petroleum industry where large amount of data exists and data-driven methods can reveal a lot of information. There is a general trend of modeling real-time data obtained through various remote sensing instruments. Mountrakis et al. (2011) summarized all important works related to remote sensing and discussed various methods and configuration through least square SVM as well as ensemble methods on how remote sensory based data can be modeled using SVM technology. Ternyik et al. (1995) used data-driven techniques in virtual measurement of pipes. Gharagheizi et al. (2017) used SVM techniques to classify to a high degree of accuracy, the conditions leading to the onset of sanding during production. The effectiveness of the low parameter requirement of least square SVM have been showcased by Ahmedi et al. (2016), where he compared the SVM methodology with other more commonly used artificial intelligence techniques for predicting water breakthrough in fractured reservoirs. Another application of least square SVM has been put forth by Mesbah et al. (2017), where they showed how an artificial intelligence technique performs better than its' peers in prediction of hydrate formation temperature for both sweet and sour gases. Ebrahimi et al. (2016) used SVM as an alternative to a reservoir simulation software in order to simulate the performance of a natural gas simulator. They were able to obtain similar degree of performance by optimizing the SVM parameters using evolutionary algorithms and continuously optimizing the evolutionary algorithm. Wang et al. (2014) developed a BBN system taking into account caprock permeability class as a surrogate for carbon dioxide leakage at saline sequestration sites due to increase in injection sites. Bassamzadeh et al. (2017) combined BBN with krigging and ANN modeling approaches in order to obtain oil production rates at sites with minimal data. They were able to improve computational speed for improved attribute selection. Bayesian Networks were used ensembled with ANN in order to survey wells with high water production prediction in gas well by Hermann et al. (2011). Li et al. (2013) showed performance of decision trees enhanced by neural networks in oil performance prediction. Aulia et al. (2014) also considered the growth of multiple decision tree and showed that Random Forests can help in strategic well test planning. Temizel et.al. (2017) studied big data analytics to enhance the oil recovery by optimizing the injection/production practices for water floods. Jia (2016) utilized machine learning based on time series analysis and neural network for control and automation of shale gas production.

Reservoir Studies and EOR

Some of the work done in the field of enhanced oil recovery and reservoir studies are presented in this section. In Saputelli et al. (2003) proposed to use hybrid models which is a combination of parametric models (neural networks etc.) and first principle models for self-learning reservoir management. He used a "model-based decision-making engine" which enables to enhance the understanding of the reservoir by using collected field data. Worthington (2005) used DDM based on neural networks and applied a dynamically-conditioned approach which fits-for-purpose cut-offs in integrated reservoir studies. Self-Organizing Network models were used in seismic analysis which helped the geophysicist to create facies maps, and wavelet transforms aid in the identification of seismic trace singularities (Lailly, 1983; Pratt, 1999). Ahmedi et al. (2016) developed a SVM based least square method to go hand-in-hand with commercial software in the simulations of recovery using chemical floods. A study was also done to use SVM to predict the effect of Nano-particles in reduction of drag effect in hydrocarbon flow in subsurface reservoirs (Di et al., 2015). A study was conducted by Ghoraishy et al. (2008) to model candidate selection as well as performance of gel treated wells using both naïve Bayesian techniques and BBN. A similar BBN was developed by Zerafat et al. (2011) for screening EOR procedures for fields. Guang-ren at al. (2008) applied SVM algorithm to predict fracture behavior and gassiness, and compared it to other AI methods. Similarly, Sujath Ali et al. (2013) showcased the application of SVM prediction capabilities for hydraulic units with a high accuracy. Prediction of oil holdup in two phase reservoirs was indicated using least square SVM, where parameters for SVM were optimized using GA.

Reservoir Management

Asset Management in the scale of the reservoir, or in other words reservoir management is one of the more innovative fields where data sciences especially artificial intelligence has been applied. The application has significantly leaned towards more popular data management tools and machine learning techniques such as Artificial Neural Networks, boosting methods and evolutionary algorithms.

Anifowose et al. (2015) developed a natural gas reservoir characterization methodology involving ensemble bootstrapping methodology to better predict reservoir parameters for better feature selection and well control improving on traditional random forest ensemble used in industry. This is a method they classify under Extreme Machine Learning. Karkevandi-Talkhooncheh (2018) developed a methods for hybrid adaptive Neuro-Fuzzy Inference system for better well placement by optimizing the existent reservoir simulator model with Particle Swarm and other evolutionary algorithm to select parameter for well placement and compared with NPV estimations and combined with generalized neuro-fuzzy algorithm. Similar study on well placement was conducted by Nwachukwu (2018) used extreme Gradient Boosting Methodology in order to simulate well placement and well to well connectivity and applying the same to multiple fluid injection cases. The model indicated a significant improvement between proxy predictions and reservoir simulation results. Pankaj et al. (2018) use data analytics for quick well completion optimization system, the methodology provides an integrated approach for big data to create proxies allowing better decision-making capabilities for operators. Yanfang et al. (2014) generated a well re-fracturing methodology for candidate well selection using neural network methodology for the Zhonyuan field.

Another area within reservoir management with several applicable areas for artificial intelligence is mature fieldrejuvination as shown by Brown et al. (2017). Brown, generated an application with his organization (QRI Analytics), which relies on rapid integration and quality control of geological and well data along with machine learning and AI technology to provide information on field development opportunities. Al-Saad et al. (2013) mention the tool AVAIL+ in their paper, which is collection of E&P data stores integrated with reservoir analytics engine with the primary focus on reservoir performance predictions. Similarly, Integrated Reservoir Management system which combines reservoir management, wellbore integrity and facilities management often use various machine learning techniques as shown by Bravo et al. (2011).

Facilities, Remediation and Management

It is possible to use advanced data-driven techniques to better understand the performance of certain oil rig equipment, rig schedules as well as remedial operations. Qu, et al. (2010) developed a data preprocessing algorithm based on SVM to better classify and predict degrees of wear in a slurry pump impeller. They combined random sub-sampling validation with SVM to identify outliers using misclassification rate. BBN technology has widely been used for the purpose of mapping feed water stream (Liu et al., 2016). A BBN model was developed by Vinnem et al. (2012) to map out the various offshore rig management strategies in order to reduce risk and prevent losses. They managed to illustrate the effect of each strategy using BBN. He et al. (2006) developed a probabilistic rule-based decision support system in order to choose the best soil remediation strategy for petroleum contaminated sites. Ceperic et al. (2016) showed that an ensemble of ANN with SVM could be used for forecasting short term Henry Hub gas prices. Fakhraver et al. (2017) used discrete BBN for performing risk analysis for possible terrorist attacks on gas pipelines and Ferreira et al. (2012) discussed how BBN could be used for vendor and supplier selection for projects. Galli et al. (1999) showed decision trees can be used for project cost evaluation processes. In 2003, Saputelli et al. proposed to use hybrid models which is a combination of parametric models and first principle models for self-learning reservoir management. He used a "model-based decision making engine", which enhanced the understanding of the reservoir by using the collected field data.

Pipelines

Data-driven techniques have been used significantly to understand the condition of pipelines under different conditions as well as risks involved with them. Lee et al. (2013) designed a system for a pipeline failure prediction system to reduce human intervention. They combined long range ultrasonic transducers with SVM algorithm based on real-time data. The system uses the ultrasonic transducers and Euclidian SVM systems to bypass the common kernel function and soft margin measures to give better realizations of failure circumstances. Liu et al. (2018) proposed an ensemble method of decision trees, which in conjunction with BBN and Genetic Algorithms can be used to plan the optimal inspection problem in pipelines. The BBN was used to simulate the corrosion growth in pipelines while the Genetic Algorithm improved the efficiency of the decision tree. Wu et al. (2015) developed a BBN model to address the cause-effect relationship to analyze the risk associated with offshore pipelines. This involved the integration of interpreted structural modeling, which was used to describe complex model, with BBN. Gu et al. (2005) generated a decision tree to analyze and classify various factors involved in causing stress, corrosion and cracking pipelines under different environmental and loading conditions.

Advantages and Disadvantages of Data-Driven Methods

Data-driven methods and computational intelligence are increasingly being used as a complementary or replacement for the physics-based models such as numerical reservoir modeling and simulation. Although, data-driven methods are effective in solving the critical problems in the oil and gas industry, there are some designing challenges that could create extra problems and could even turn the model into garbage. Mainly, there are some focal points that need to be described clearly such as optimum number of layers to solve the problem, number of units needed for each layer, the generalization ability of the algorithm and the boundaries of a training set to handle the problem successfully. When the data is applied blindly especially in a situation, where sufficient amount of data about the problem does not exist or when the modeled system is not stable during the period covered by the model, algorithmic bias could become a risk (Solomatine et al., 2008).Table 1, provides a summary of some of the advantages, disadvantages and areas of application for the data-driven methods discussed in this paper.

In most of these methods, the link between raw data and the generated knowledge is hidden so transforming data into valuable insights is a challenge (Hartman et al., 2017). Moreover, since nature is

fuzzy, it is usually hard to implement only one data-driven method to capture the behavior of the whole system. To overcome these issues, the contemporary trend suggests using hybrid models, a combination of different data- or physics-driven methods, to generate a single solution.

One of the major factors in data analytics is the rise of Big Data (BD), which is becoming the center of attention worldwide. BD is characterized by certain salient features and its surrounding complications. According to Fan, et al. (2014), the major challenges surrounding the rise of BD are as follows:

- High dimensionality of data leading to erroneous correlation, noise, senseless clustering and "incidental" homogeneity.
- High computational time and instability in the software and high infrastructural cost relating to these computational expectations and storage.
- Issues of heterogeneity and experimental variation because of the biases due to multiple distributed sources of data matrices.

Table 1—Applications, advantages and disadvantages of data-driven methods

Data-driven Technique	Application	Advantage	Disadvantage
Linear Regression	Quick and simple prediction of linear data	• Simple • Computational Inexpensive	• Inaccurate in some cases • Limited attribute representation
Decision Trees	Classification and regression analysis and especially useful in simplified description.	• Performs well despite missing data and various kinds of data formats • Low effect of outliers • Provided pruning techniques, decision trees are pretty robust in presence of noise	• High number of attribute can have adverse effect on the performance of the tree • Simple Decision tree boundaries are rectilinear • Prone to overfitting of data
Support Vector Machines	Non-linear and linear data set with noisy data; used for high accuracy regression.	• High Accuracy • Robust even in presence of data inaccuracies • Fast Evaluation of learned target function	• Training is computationally Intensive • Intensive domain specific implementation procedure
Artificial Neural Networks	Highly versatile and useful in parameter optimization, regression and classification.	• Highly transmutable • Robust performance despite noise • Algorithm is highly durable and can maintain performance despite failure in specific layer	• Difficulty in interpretations of procedure of obtaining results • Training data is computationally intensive
Fuzzy Logic	Used when data are not linearly separable.	• Introduces degrees of membership or classification (degrees of truth) to various other techniques	• Makes the other techniques comparatively intensive in terms of computation
Genetic Algorithms	Used for optimization, parameter selection and when data is not uniform.	• High accuracy in parameter selection • Low memory usage • Robust despite outliers	• Computationally intensive • Black-box • Probability of premature solution convergence
Bayesian Belief Networks	Great for finding interrelation between parameters and cause-effect relations.	• Casual Relationships can be explored • Facilitates easy use of prior knowledge • Prevents over-fitting	• Computationally intensive • Poor performance on small datasets

Further to these issues, the reader is referred to the work of Sivarajah, et al. (2017), who compiles the work of over 200 research studies across several fields explaining data challenges relating to characteristics of the data itself, data processing problems and management issues such as privacy, security and ethics (Fig. 9).

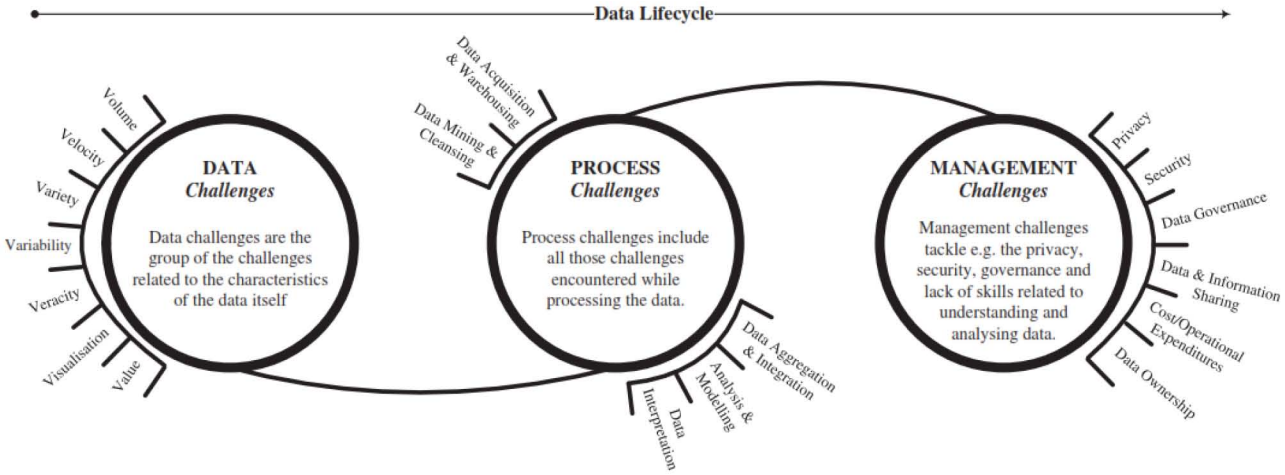

Figure 9—Conceptual Classification of Data Challenges (Sivarajah et al., 2017)

Acceptance of Data-Driven Methods in the Oil & Industry

The oil & gas industry is one of the major sensor-based industries with a lot of data collecting sensors that are installed downhole and on the surface. Companies also monitor their assets closely to calculate their reservoir production as well as to predict the future performance of their reservoirs. As a result, the petroleum industry has to deal with considerably large volumes of structured and unstructured data from various sources. Data-driven models are increasingly being used in the industry to analyze such kinds of data, in particular finding connections between the input and output state variables without explicit information about the physical behavior of the system. More and more companies are realizing that they can utilize this available data to better optimize their overall performance in different areas, such as increasing the production capacity of the reservoir, forecasting extreme events, or simulating fluid flow. Despite numerous advantages that data-driven techniques can provide the industry, many oil and gas companies are still reluctant to adopt these technologies. Some possible reasons associated with this reluctance could be:

1. It is a challenge to convince engineers that data is a reality driven: Senior executives seem to be aware of the importance of the big data and data analytics, however, team members such as petroleum engineers usually have a hard time to define a path to value the data-driven methods which are not part of any petroleum engineering curriculum at the universities (Mohaghegh et al., 2017).
2. Lack of cross-functional ownership: Developing the capability to build advanced analytics models is a C-level agenda item for the senior management team. The task is usually delegated to CIO or CTO as a technology project, however it requires cross-functional ownership and participation.
3. Cost effectiveness: Replacing or maintaining sensors such as downhole temperature or pressure sensors to generate the data required for data-driven analytics could be expensive.
4. Threats of cyber-attacks: As the digital oil field becomes a reality the threat of external cyberattacks increase. In fact, in 2016, the O&G production operation ranked highest in cyber vulnerability in upstream operations. Therefore, oil companies should have a cybersecurity expert to keep their assets safe (Ponemon I. LLC et al., 2017; Mittal et al., 2017). Based on the prediction of cybersecurity ventures, there will be 3.5 million unfilled cybersecurity jobs by 2021 (Morgan at al., 2016).
5. Robustness of data: Geoscientists and engineers usually practice in offices with totally different environments compared to the field. In these environments, either onshore or offshore, operators need to give quick and safe decisions based on their past experiences. Therefore, the data driven technology must meet standards of robustness and reliability in order to be accepted and applied by operators.

Conclusion

Quantification of uncertainty, minimizing risk and maximizing profit as well as the speed are the key elements in decision making in the O&G industry. The proliferating amount of data constantly generated, thanks to the advances in the technology, has the potential to dramatically enhance the intuitive decisions made in various day to day activities. However, the prospective benefits of the data are only achievable if the right tools are employed to integrate different types of data and transform it in to useful information that help in deriving smart conclusions. This study provided a concise yet comprehensive overview of the most common data-driven techniques that can be applied in different disciplines in the O&G industry. Advantages and disadvantages of each of these methodologies along with several examples of successful applications of data-driven techniques were also provided. Finally, the current acceptance status of the data-driven methods in the industry was discussed and some of the obstacles delaying the full recognition of their importance were provided.

References

1. Saputelli, L., Nikolaou, M., & Economides, M. J. (2003, January). Self-learning reservoir management. In SPE Annual Technical Conference and Exhibition. *Society of Petroleum Engineers*.
2. Mishra, S., Datta-Gupta, A., 2017. *Applied Statistical Modeling and Data Analytics A Practical Guide for the Petroleum Geosciences*, Elsevier
3. Holdaway, K.R., 2014. *Harnessing Oil and Gas Big Data With Analytics*. John Wiley & Sons, Hoboken, NJ.
4. Saputelli, L., 2016. Technology focus: petroleum data analytics. *Soc. Pet. Eng.* ttps://doi.org/10.2118/1016-0066-JPT.
5. Haan, C.T., 1986. *Statistical Methods in Hydrology*. Iowa University Press, Ames. 376.
6. Hastie, T., Tibshirani, R., Friedman, J.H., 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
7. Vapnik, V. N., Drucker, H., Burges, J. C., Kaufman, L., Smola, A. J., (1997); "Support Vector Regression Machines", *in Advances in Neural Information Processing Systems 9, NIPS* 1996, 155–161, MIT Press.
8. Alisneaky, 2011. https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Kernel_Machine.png
9. Zadeh, L. A. (1965). Information and control. *Fuzzy sets*, **8**(3), 338–353
10. Sivanandam, S. N., Deepa, S. N. 2011. *Principles of Soft Computing*, 2nd Edition. Wiley India, New Delhi, India.
11. Bian, X. Q., Han, B., Du, Z. M., Jaubert, J. N., Li, M. J. Integrating support vector regression with genetic algorithm for CO2-oil minimum miscibility pressure (MMP) in pure and impure CO2 streams. *Fuel* **182** (2016) 550–557
12. Bhandari, J., Abbassi, R., Garaniya, V., Khan, F. Risk Analysis of Deepwater Drilling Operations using Bayesian Network. *Journal of Loss Prevention in the Process Industry* **38** (2015) 11–23.
13. Taner, M. T., Koehler, F., & Sheriff, R. E. (1979). Complex seismic trace analysis. *Geophysics*, **44**(6), 1041–1063.
14. Ouenes, A. (2000). Practical application of fuzzy logic and neural networks to fractured reservoir characterization. *Computers & Geosciences*, **26**(8), 953–962.
15. Zhang, L., Wu, S., Zheng, W., Fan, J. A Dynamic and Quantitative Risk Assessment method with Uncertainties for Offshore Managed Pressure Drilling Phases. *Safety Sciences* **104** (2018) 39–54.

16. Al-Anazi, A. F., Gates, I. D. Support Vector Regression to Predict Porosity and Permeability: Effect of Sample Size. *Computers and geoscience* **39** (2012) 64–76.
17. Behnoud Far, P., Hosseini, P., Azizi, A. Permeability Determination of cores based on their Apparent Attributes in the Persian Gulf region using Naïve Bayesian and Random Forest Algorithm. *Journal of Natural Gas Science and Engineering* **37** (2017) 52–68.
18. Ozkaya, S. I. Using Probabilistic Decision Trees to Detect Fracture Corridors from Dynamic Data in Mature Oil Fields. SPE Middle East Oil and Gas Show and Conference, Bahrain. 2006. SPE 105015.
19. Chamkalani, A., Zendehboudi, S., Chamkalani, R., Lohi, A., Elkamel, A., Chatzis, I. Utilization of Support Vector Machine to Calculate Gas Compressibility Factor. *Fluid Phase Equilibria* **358** (2013) 189–202.
20. El-Sebakhy, E. A. Forecasting PVT Properties of Crude Oil Systems based on Support Vector Machines Modeling Scheme. *Journal of Petroleum Science and Engineering* **64** (2009) 25–34.
21. Tohidi-Hosseini, S. M., Hajirezanie, S., Doulatabadi, M. H., Sarapardeh, A. H., Mohammadi, A. H. Toward Predicttion of Petroleum Reservoir Fluid Properties: A Rigorous Model for Estimation of Solution Gas-Oil Ratio. *Journal of Natural Gas Science and Engineering* **29** (2016) 506–516.
22. Ahmadi, M. A., Mahmoudi, B. Development of Robust Model to Estimate Gas-Oil Interfacial Tension using Least Square Support Vector Machine: Experimental and Modeling Study. *Journal of Supercritical Fluids* **107** (2016) 122–128.
23. Cranganu, C., Breaban, M. Using Support Vector Regression to Estimate Sonic Log Distributions: A case study from the Anadarko Basin, Oklahoma. *Journal of Petroleum Science and Engineering* **103** (2013) 1–13.
24. Akande, K. O., Owolabi, T. O., Olatunji, S. O., Abdulraheem, A. A hybrid Particle Swarm optimization and Support Vector Regression model for modelling Permeability Prediction of Hydrocarbon Reservoir. *Journal of Petroleum Science and Engineering* **150** (2017) 43–53.
25. Tan, M., Song, X., Yang, X., Wu, Q. Support Vector Regression Machine technology for Total Organic Carbon Content Prediction from wireline logs in Organic Shale: A comparative Study. *Journal of Petroleum Science and Engineering* **26** (2015) 792–802.
26. Masoudi, P., Tokhmechi, B., Jafari, M. A., Zamanzadeh, S. M., Sherkati, S. Application of Bayesian in determining Productive Zones by Well Log data in Oil wells. *Journal of Petroleum Science and Engineering* **94-95** (2012) 47–54.
27. Anifowose, F., Labadin, J., Abdulraheem, A. Improving the prediction of Petroleum Reservoir Characterization with a Stacked Generalization Ensemble Model for Support Vector Machines. *Applied Soft Computing* **26** (2015) 483–496.
28. Ahmadi, M. A. Towards Reliable Model for Prediction Drilling Fluid Density at Wellbore Conditions: A LSSVM model. *Neurocomputing* **211** (2016) 143–149.
29. Fatehi, M., Asadi, H. H. Data Integration Modeling applied to Drill Hole Planning through Semi-Supervised Learning: A case study from the Dalli Cu-Au porphyry Deposit in Central Iran. *Journal of African Sciences* **128** (2017) 147–160.
30. Zhang, L., Wu, S., Zheng, W., Fan, J. A Dynamic and Quantitative Risk Assessment method with Uncertainties for Offshore Managed Pressure Drilling Phases. *Safety Sciences* **104** (2018) 39–54.
31. Al-Yami, A. S., Al-Shaarari, A., Al-Bahrani, H., Wagle, V. B., Al-Gharbi, S., Al-Khudiri, M. B. Using Bayesian Network to Develop Drilling Expert Systems. SPE Heavy Oil Conference and Exhibition, Kuwait City, Kuwait. 2016. SPE-184168-MS.
32. Sule, I., Khan, F., Butt, S., Yang, M. Kick Control Reliability Analysis of Managed Pressure Drilling Operation. *Journal of Loss Prevention in the Process Industry* **52** (2018) 7–20

33. Chang, Y., Chen, G., Wu, X., Ye, X., Ye, J., Chen, B., Xu, L. Failure Probability Analysis for Emergency Disconnect of Deepwater Drilling Riser using Bayesian Network. *Journal of Loss Prevention in the Process Industry* **51** (2018) 42–53.
34. Cai, B., Liu, Y., Liu, Z., Tian, X., Dong, X., Yu, S. Using Bayesian Networks in Reliability Evaluation for Subsea Blowout Preventer Control Systems. *Reliability Engineering and System Safety* (2012) 32–41.
35. Kormaksson, M., Vieira, M.R., Zadrozny, B., 2015. A Data Driven Method For Sweet Spot Identification In Shale Plays Using Well Log Data, SPE Digital Energy Conference and Exhibition, The Woodlands, Texas, USA, SPE-173455-MS.
36. Bakshi, A., Uniacke, E., Korjani, M., Ershaghi, I. 2017.A Novel Adaptive Non-Linear Regression Method to Predict Shale Oil Well Performance Based on Well Completions and Fracturing Data, SPE Western Regional Meeting, Bakersfield, California, SPE-185695-MS.
37. Temizel, C., Aktas, S., Kirmaci, H., Susuz, O., Zhu, Y., Balaji, K., Ranjith, R., Tahir, S., Aminzadeh, F., Yegin, C. 2016. Turning Data into Knowledge: Data-Driven Surveillance and Optimization in Mature Fields, SPE Annual Technical Conference and Exhibition, Dubai, UAE, SPE-181881-MS.
38. Mountrakis, G., Im, J., Ogole, C. Support Vector Machines in Remote Sensing; A Review. *ISPRS Journal of Photogrammetry and Remote Sensing* **66** (2011) 247–259.
39. Ternyik IV, J., Bilgesu, H. I., Mohaghegh, S., & Rose, D. M. (1995, January). Virtual measurement in pipes: Part 1-Flowing bottom hole pressure under multi-phase flow and inclined wellbore conditions. In SPE Eastern Regional Meeting. *Society of Petroleum Engineers*.
40. Gharagheizi, F., Mohammadi, A. H., Arabloo, M., Shokrollahi, A. Prediction of Sand Production onset in Petroleum Reservoirs using a Reliable Classification approach. *Petroleum* **3** (2017) 280–285.
41. Ahmadi, M. A., Ebadi, M., Hosseini, S. M. Prediction of Breakthrough Time of Water Coning in the Fractured reservoirs by implementing low parameter Support Vector Machine approach. *Fuel* (2016) 579–589.
42. Mesbah, M., Soroush, E., Rezakazemi, M. Development of a Least Squares Support Vector Machine model for Prediction of Natural Gas Hydrate Formation Temperature. *Chinese Journal of Chemical Engineering* **25** (2017) 1238–1248.
43. Ebrahimi, A., Khamehchi, E. Developing a Novel Workflow for Natural Gas Lift Optimization using advanced Support Vector Machine. *Journal of Natural Gas Science and Engineering* **28** (2016) 626–638.
44. Wang, Z., Small, M. J., A Bayesian approach to CO2 leakage detection at Saline Sequestration Sites using Pressure Measurements. *International Journal of Greenhouse Gas Control* **20** (2014) 188–196.
45. Bassamzadeh, N., Ghanem, R. Probabilistic Data-Driven Prediction of Wellbore Signatures in High Dimensional Data Using Bayesian Networks. *SPE Journal* 2018.
46. Hermann, R., Ratanavani, S., Mcholvilert, S. L., Nitura, J., Chandakaew, R., Jiraratchwaro, C., Vitoonkijvanich, S., Sarisittitham, S. Water Production Surveillance workflow using Neural Network and Bayesian Network Technology: A Case Study for Bongkot North Field, Thailand. IPTC, Bangkok, Thailand. 2012. IPTC 15015.
47. Li, X., Chan, C. W., Nguyen, H. H. Application of the Neural Decision Tree approach for Prediction of Petroleum Production. *Journal of Petroleum Science and Engineering* **104** (2013) 11–16.
48. Aulia, A., Rahman, A., Velasco, J. J. Q. Strategic Well Test Planning using Random Forest. SPE Intelligent Energy Conference and Exhibition, Utrecht, Netherlands. 2014.

49. Temizel, C.; Nabizadeh, M.; Kadkhodaei, N.; Ranjith, R., Suhag, A., Balaji, K.; Dhannoon, D., 2017, Data-Driven Optimization of Injection/Production in Waterflood Operations, SPE Intelligent Oil and Gas Symposium, Abu Dhabi, UAE, SPE-187468-MS.
50. Jia, X.; Zhang, F., 2016. Applying Data-Driven Method to Production Decline Analysis and Forecasting, SPE Annual Technical Conference and Exhibition, Dubai, UAE, SPE-181616-MS.
51. Worthington, P. F. 2005. The Application of Cutoffs in Integrated Reservoir Studies. In SPE Annual Technical Conference and Exhibition. *Society of Petroleum Engineers*.
52. Lailly, P. 1983. *The seismic inverse problem as a sequence of before stack migrations*.
53. Pratt, R. G. 1999. Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model. *Geophysics*, **64**(3), 888–901.
54. Ahmadi, M. A., Purnik, M. A predictive model of Chemical Flooding for Enhanced Oil Recovery purposes: Application of Least Square Support Vector Machine. *Petroleum* **2** (2016) 177–182.
55. Di, Q. F., Hua, S., Ding, W. P., Gong, W., Cheng, Y. C., Ye, F. Application of Support Vector Machine in Drag Reduction Effect Prediction of Nanoparticles Adsorption method on Oil Reservoir's Micro-channels. *Journal of Hydrodynamics* 2015, **27**(1): 99–104.
56. Ghoraishy, S. M., Liang, J. T., Green, D. W., Liang, H. Application of Bayesian Networks for predicting the Performance of Gel-Treated wells in the Arbuckle Formation, Kansas. SPE/DOE Improved Oil Recovery Symposium, Tulsa, Oklahoma. 2008. SPE 113401.
57. Zerafat, M. M., Ayatollahi, S., Mehranbod, N., Barzegari, D. Bayesian Network Analysis as a Tool for Efficient EOR Screening. SPE Enhanced Oil Recovery Symposium, Kuala Lumpur, Malaysia. 2011. SPE 143282.
58. Guang-ren, S. H. I.Superiorities of Support Vector Machine in Fracture Prediction and Gassiness Evaluation. *Petroleum Exploration and Development* **25**; 5(2008) 588–594.
59. Ali, S. S., Nizamuddin, S., Abdulraheem, A., Hassan, M. R., Hossain, M. E. Hydraulic Unit Prediction using Support Vector Machine. *Journal of Petroleum Science and Engineering* **110** (2013) 243–252.
60. Anifowose, F. A., Labadin, J., Abdulraheem, A. Ensemble model of Non-Linear Feature Selection-based Extreme Learning Machine for Improved Natural Gas Reservoir Characterization. *Journal of Natural Gas Science and Engineering* **26** (2015) 1561–1572
61. Karkevandi-Talkhooncheh, A., Sharifi, M., Ahmadi, M. Application of Hybrid Adaptive Neuro-Fuzzy Inference System in Well Placement Optimization. *Journal of Petroleum Science and Engineering* **166** (2018) 924–947
62. Nwachukwu, A., Jeong, H., Pyrcz, M., Lake, L. W. Fast Evaluation of Well Placements in Heterogenous Reservoir Models using Machine Learning. *Journal of Petroleum Science and Engineering* **163** (2018) 463–475
63. Pankaj, P., Geetan, S., MacDonald, R. Need for Speed: Data Analytics Coupled to Reservoir Characterization Fast Tracks Well Completion Optimization. SPE Canada Unconventional Resources Conference, Calgary, Alberta. March 2018. SPE-189790-MS
64. Yanfang, W., Salehi, S. Refracture Candidate Selection using Hybrid Simulation with Neural Network and Data Analysis Techniques. *Journal of Petroleum Science and Engineering* **123** (2014) 138–146
65. Brown, J. B., Salehi, A., Benhallam, W., Matringe, S. F. Using Data-Driven Technologies to Accelerate the Field Development Planning Process for Mature Field Rejuvenation. SPE WRM, Bakersfield, California. April 2017. SPE-185751-MS
66. Al-Saad, B., Murray, P. A., Vanderhaeghen, M., Yannimaras, D., Naime, R. K. Development and Implementation of the AVAILS+ Collaborative Forecasting Tool for Production Assurance in

the Kuwait Oil Company, North Kuwait. SPE Kuwait Oil & Gas Show and Conference. October 2013. SPE 167375

67. Bravo, C., Saputelli, L., Castro, J. A., Rios, A., Rivas, F., Aguilar-Martin, J. Automation of the Oilfield Asset via an Artificial Intelligence (AI)-Based Integrated Production Management Architecture (IPMA). SPE Digital Energy Conference & Exhibition, Woodlands, Texas. April 2011. SPE 144334.
68. Qu, J., Zuo, M. J. Support Vector Machine based Data Processing Algorithm for Wear Degree Classification of Slurry Pump Systems. *Measurement* **43** (2010) 781–791.
69. Liu, Z., Liu, Y., Wu, X., Yang, D., Cai, B., Zheng, C. Reliability Evaluation of Auxiliary Feedwater System by Mapping GO-FLOW models into Bayesian Networks. *ISA Transactions* **64** (2016) 174–183.
70. Vinnem, J. E., Bye, R., Gran, B. A., Kongsvik, T., Nyheim, O. M., Okstad, E. H., Seljelid, J., Vatn, J. Risk Modeling of Maintenance Work on Major Process Equipment on Offshore Petroleum Installations. *Journal of Loss Prevention in the Process Industries* **25** (2012) 274–292
71. He, L., Chan, C. W., Huang, G. H., Zeng, G. M. A Probabilistic Reasoning-based Decision Support System for Selection of remediation Technologies for Petroleum-contaminated Sites. *Expert Systems with Applications* **30** (2006) 783–795.
72. Ceperic, E., Zikovic, S., Ceperic, V. Short-term Forecasting of Natural Gas Prices using Machine Learning and Feature Selection Algorithm. *Energy* **140** (2017) 893–900.
73. Fakhravar, D., Khakzad, N., Reniers, G., Cozzani, V. Security Vulnerability Assessment of gas Pipelines using Discrete-time Bayesian Network. *Process Safety and Environmental Protection* **3** (2017) 714–725.
74. Ferreira, L., Borenstein, D. A fuzzy Bayesian model for Supplier Selection. *Expert Systems with Applications* **39** (2012) 7834–7844.
75. Galli, A., Armstrong, M., Jehl, B. Comparing Three Methods for Evaluating Oil Projects: Option Pricing, Decision Trees and Monte Carlo Simulations. SPE Hydrocarbon Economics and Evaluation Symposium, Dallas, Texas. 1999.
76. Lee, L. H., Rajkumar, R., Lo, L. H., Wan, C. H., Isa, D. Oil and Gas Pipeline Failure Prediction System using Long Range Ultrasonic Transducers and Euclidean-Support Vector Machines Classification Approach. *Expert Systems with Applications* **40** (2013) 1925–1934
77. Liu, X., Zheng, J., Fu, J., Nie, Z., Chen, G. Optimal Inspection Planning of Corroded pipelines using BN and GA. *Journal of Petroleum Science and Engineering* **163** (2018) 546–555.
78. Wu, W. S., Yang, C. F., Chang, J.C., Chateau, P. A., Chang, Y. C. Risk Assessment by Integrating Interpretive Structural Modeling and Bayesian Network, case of Offshore Pipeline Project. *Reliability Engineering and System Safety* **142** (2015) 515–524.
79. Gu, B., Kania, R., Gao, M., Fiel, W. Development of SCC Susceptibility Model using Decision Tree Approach. NACE International Corrosion Conference, Houston, Texas (2005). Paper No. 05479
80. Solomatine D., See L.M., Abrahart R.J. 2008. *Practical Hydroinformatics Computational Intelligence and Technological Developments in Water Applications*, Springer.
81. Hartmann, T., Moawad, A., Fouquet, F., Nain, G., Klein, J., Traon, Y. L., Jezequel, J. M. 2017. *Model-Driven Analytics: Connecting Data, Domain Knowledge, and Learning*.
82. Fan, J., Han, F., Liu, H. (2013). Challenges of Big Data Analysis. *National Science Review*, **1**: 293–314.
83. Sivarajah, U., Kamal, M. M.,Irani, Z., Weerakkody, V. Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Buisness research* **70**(2017) 263–286.
84. Mohaghegh S.D. 2017. *Shale Analytics*, Shahab D. Springer International Publishing.

85. Ponemon I. LLC et al. 2017. "The state of cybersecurity in the oil & gas industry: United States," https://ics-cert.us-cert.gov/#monitornewsletters
86. Mittal A., Slaughter A., Zonneveld P. 2017. *Protecting the connected barrels. Cyber Security for Upstream Oil and Gas*. https://www2.deloitte.com/insights/us/en/industry/oil-and-gas/cybersecurity-in-oil-and-gas-upstream-sector.html#endnote-3
87. Morgan et al. 2016. https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016