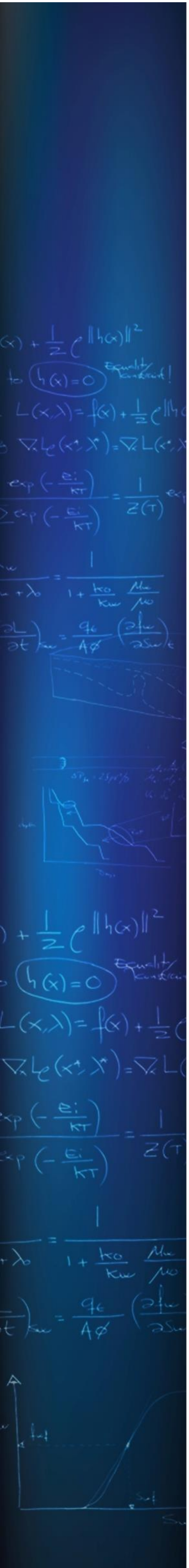

TRiCAM Project Proposal
Institute for Applied Computational Science
Harvard University

THE IMITATION GAME FOR PETROLEUM ENGINEERS

UNDERSTANDING THE BEHAVIOR OF OIL WELLS THROUGH DEEP LEARNING

Quantum Reservoir Impact

May 2015

Kun Liu, PhD
Sébastien Matringe, PhD

QRI

PROBLEM STATEMENT

Scientific Context

Understanding intelligence has been one of the great questions of humanity. Religion, philosophy, art, science; the problem has been approached from all angles but remains unsolved. Since the invention of the programmable computer, many have started to believe that Artificial Intelligence (AI) was achievable. The field of AI was born and has been in continuous evolution since, alternating between periods of fast breakthroughs and relative stagnation. Today, AI is thriving with many practical applications in labor automation, speech and object recognition, high-frequency finance, or medical diagnosis, just to name a few.

The key challenge for AI systems is to match the robustness and efficiency of human intelligence in complex, real-world applications. In the early days of machine learning, the field focused on challenges that were ideally suited for computers to solve: problems that could be described programmatically by a list of formulae and mathematical rules. A typical example is chess, and the famous victory of IBM's Deep Blue over Garry Kasparov in 1997. More recently, scientists have been extending the capabilities of machine learning to tackle problems that are intuitive for humans but hard to describe formulaically.

The ultimate intelligence test was proposed by Alan Turing in his seminal 1950 article "Computing Machinery and Intelligence". The Imitation Game is used to define artificial intelligence. Simply put, if a human examiner is not able, through interrogation, to distinguish between a machine and a human, then the machine should be considered intelligent.

QRI proposes to create a tool able to intelligently describe the behavior of oil wells. The production profiles of these wells can be complex, driven by reservoir physics, marked by a variety of operational events and obfuscated by data noise. Petroleum Engineers are usually able to decipher the well production profiles, understand the underlying behavior, forecast the expected production and identify opportunities for performance improvements. This investigation process is however fairly time-consuming, usually taking a few hours per well. For fields with hundreds or thousands of wells, the work cannot be done as routinely as it should, which quickly creates a vast portfolio of unidentified opportunities for any mature field.

Over the past few years, QRI has been working on formalizing this analytic process. We have developed technology to automate or assist part of the diagnosis and forecast, using advanced applied mathematics and engineering techniques. The QRI algorithms implement the problem-solving approach of experienced engineers and geoscientists. The technology can thus hardly be described as artificial intelligence and is better classified as "hard-coded" intelligence.

The recent developments in machine learning technologies, namely deep-learning algorithms, offer a glimpse of hope. These new AI techniques are designed to better mimic the actual human learning process and have brought tremendous advances in pattern and object recognition.

The question we would like to answer this summer is the following: "Can deep-learning algorithms be used to understand the behavior of oil wells?"

Deep Learning

Deep learning is the most popular new development in machine learning technologies. Advances in Neuroscience continuously improve our understanding of how the biological cortex works. In the early 2000's, the improved description of cortical learning started to highlight differences between how the classical artificial neural networks (ANN) of the 1980's were designed and trained and how actual neurons are organized and triggered. This paradigm-shift, popularized in the book On Intelligence by Jeff Hawkins, led to a revival of machine learning technology, which marked the end of the latest "AI winter".

Deep learning algorithms are a form of ANN. They take their names from their network architecture. Hornik's Universal Approximation Theorem (1991) showed that a feed-forward neural network with a single hidden layer could approximate any continuous function in any compact multidimensional real domain with a finite number of nodes. This means that a very simple ANN architecture can theoretically be used to approximate most functions encountered in real life. What the theorem doesn't mention is that for many problems, such an architecture is far from optimal and a solution can be found much faster and with much less resources by adding depth to the ANN. Deep networks rely on more hidden layers to achieve better results.

The learning algorithms used in deep learning is also novel. Rather than the standard back-propagation algorithms, used in classical ANN, deep learning start with an unsupervised learning phase, where the network learns fundamental features in the dataset. Supervised learning is performed at the last stage of the algorithm, when salient features of the dataset have been identified.

To take an example, we can analyze how a child learns to recognize a cat from other animals or objects. The classical understanding was that the child would be exposed to cat images and sightings and over time, his parents would indicate to him that what he was seeing was in fact called a cat. This is supervised learning. The brain is exposed to an example (an image of a cat) and is supplied with the solution (this is or is not a cat). Over time, the brain learns to make the right association. The current understanding of human learning has evolved. Now, researchers recognize two phases in the learning process. First, the child will be exposed to cats. He will independently start to form the concept of a cat in his head and eventually associate a cat with its features – a tail, four legs, a round face, pointy ears, big eyes, retractable claws, etc. When the concept is well formed, he will ask around him what to name the concept that he's formed and it's only at this time when the supervised learning occurs. Prior to that, the learning was unsupervised. Deep learning algorithms rely on this approach.

This example was implemented by Google [x] in 2012. A network of 16,000 processors and one billion connections was assembled and was exposed to 10 million randomly selected and unlabeled YouTube video thumbnails over the course of three days. After being presented with a list of 20,000 different items, the network started to recognize pictures of cats: some of the neurons learned to respond strongly to pictures of cats. This was a remarkable breakthrough in machine learning, the network learned the concept of cat by itself, without being supervised to do so. Today, deep learning has been successfully applied to other visual recognition applications, as well as to other fields, such as speech recognition, medical diagnosis, and time-series forecasting where it's used to extract useful features and patterns from complicated systems.

Application to the Petroleum Industry

Wells drilled in petroleum reservoirs produce a mixture of oil, gas and water. These wells can be exploited for decades, the oldest well in the U.S. being 153 years old (McClintock No. 1). Over their lifetime, the behavior of these wells will evolve. When first drilled, most wells produce oil at a very high rate. As the oil flows up out of the well, from the reservoir to the surface, hydrocarbon gas is usually released from the pressurized liquid hydrocarbons, in a similar fashion as carbon dioxide evolves from soda water when a bottle is opened. This gas is produced at the surface along with the oil. As the reservoir gets depleted the oil production rate will start declining fairly naturally. Eventually, the well will start to produce some amount of water as aquifers start to encroach on the deflating reservoir.

Production metrics are often used to analyze the level of maturity and health of a well. The most simple metrics, the gas-oil ratio (GOR) and water-cut (WCT), represent the proportion of the volume of gas to oil and water to total liquid respectively. Mature wells in Texas or Oklahoma often produce at a water-cut above 95%. The trend in the water-cut can be used to understand whether the water being produced comes from below the reservoir or from its sides. An increasing GOR indicates that the pressure of the reservoir is dropping below the "bubble-point" pressure and that the gas is starting to evolve from the oil directly in the reservoir.

During their lifetime, wells are also modified through "work-overs". These repairs, upgrades or modifications are designed to add more oil production or reduce the amount of unwanted fluids (water and gas). Pumps can be installed, production pipelines can be changed, valves can be opened or closed, the well itself can be made to produce from a different reservoir formation, etc. These production events are usually detectable in the production history of a well: drastic changes in the production behavior usually indicate that something has been physically modified in the well. Sometimes, records of these modifications are well kept, but often the well history is lost or was never recorded.

Reservoir conditions also affect the well behavior. Following the primary depletion of a reservoir, a pressure maintenance strategy is usually developed, where fluids (often water) are injected into the reservoir in order to compensate for the fluids produced. If the reservoir pressure increases, the decline of the well production rate slows down and can even be reversed so the well production starts to increase again.

To further complexify the problem, noise can be significant in the production data. Often the company developing the field measures accurately and continuously the total oil production rate for the field, since it is the major revenue source. However, water volumes are not necessarily recorded accurately. Gas rates can be even worse and they are often more difficult to record since the produced gas is sometimes just flared (this was common practice historically in the U.S. and is still done in some parts of the world). Even the oil rate of each individual well is not always perfectly recorded. Routine rate measurements by well are performed and used to back-allocate the total field production.

The production rate of oil wells, although mundane in appearance, is the convolution of reservoir effects, wells actions and potential measurement errors. Petroleum Engineers constantly have to analyze well production behavior in order to extract the wealth of information hidden in this seemingly simple multivariate time-series. Trained engineers can quickly understand trends in the production history of wells and gain insights as to what might have happened and how to improve the current performance.

Physics-based models are often created to complement the engineer's trained eyes. These models can get fairly complex. They usually take time to build and often require data that is not readily available. Some of the necessary data might have never been made available to the current operator (following the acquisition of the field from another company for example), some of the data might have been destroyed in wars, or be in paper files in a basement across the world, or might simply have never been recorded, which is often the case for wells drilled in the early 20th century. Physics-based models often require a clean dataset to be useful.

Data-driven analytics attempts to approach the problem from a difference angle: extract as much information from a dataset such as trends and patterns as possible. Erroneous data is often detected and discarded and the missing data is inferred robustly. Such methods have a tremendous future in the oil and gas industry. Many companies are heavily involved in the development of such technologies. QRI, for one, has been a pioneer in the field and is enjoying great success in applying "big data" analytics to a variety of oil fields around the world.

Most of the data-driven analytics currently used in the oil industry leverages classical statistical or machine learning algorithms. Although it has been applied to time-series analysis in Finance or other industries, deep learning has not yet been used in the oil industry and QRI expects its first application to provide tremendous insights into well behaviors. There are currently over 1.1 million active wells in the U.S., which represent less than 10% of the world's oil production. An improvement in how these wells are analyzed and diagnosed could lead to significant opportunities throughout the industry.

PROJECT DESCRIPTION

Project Objectives

The general objective of the project is to develop a machine learning tool and train it to understand the behavior of oil wells. More specifically, the tool should be able to identify historical production events and to forecast the expected production trends of a well under various scenarios. This tool would therefore be an artificial intelligence system designed to extract information from a multivariate time-series.

Recommendations

To do so, QRI recommends the following machine-learning process:

1. **Data pre-processing.** The dataset provided contains a variety of wells from conventional fields throughout the world. The dataset contains the monthly oil, water and gas production rates of each well over time. The dataset is delivered in a clear and useable format, but also contains natural noise and variations. Some pre-processing is most likely necessary to improve the performance of the machine learning algorithms in the later steps. Data exploration, visualization, clustering, normalization and reduction could all be useful and healthy for pre-processing steps to take.
2. **Unsupervised Feature Learning.** This step is the key to a deep-learning algorithm. The fundamental idea behind deep learning is that machine learning algorithms perform better when exposed to higher-level learned features rather than raw high-dimensional data. For that reason, a first step of unsupervised feature learning designed to extract fundamental features in the raw data could be used as a pre-processor to a more standard supervised machine learning method.
3. **Supervised Machine Learning.** Finally, a machine learning algorithm could be applied on learned features to understand the well behaviors and forecast the production in time.

A specific “ready-to-use” solution that could be investigated is the Hierarchical Temporal Memory (HTM) learning algorithm developed by Numenta and available to the public as NuPIC. QRI has not tested the HTM yet, but feels that it offers promises, although it was really developed for streaming data and should therefore probably be adapted slightly to work on this project.

In general, QRI recommends the team to use a simple programming language with libraries publically available. Internally, QRI relies heavily on Matlab, but R and Python are well known solid alternatives, each with their own advantages and drawbacks.

DATASET DESCRIPTION

The dataset provided is extracted from two super-giant fields, containing a total of seven reservoirs (three in one field and four in the other). Unfortunately, QRI cannot reveal the name of the client, fields or reservoirs used in this study for confidentiality reasons.

The dataset has been modified in two ways:

- For confidentiality reasons, the whole production has been multiplied by a single, constant factor. This is to respect our client's request to maintain the privacy of the data. This should of course not affect the algorithms at all.
- No further modification of the data was made, but QRI has kept a small percentage of wells (10% or less) and the most recent production data to perform a blind test at the end of the project. As always in machine learning, we recommend the team to segment their dataset into a training, validation and testing sets in order to quantify the learning quality and to avoid over-training.

The dataset consists of well by well historical production data aggregated in an Excel spreadsheet. The Excel file contains seven sheets, each of which corresponds to the dataset of an individual reservoir. The name of the sheet is the abbreviation of the reservoir name. For example, BEAT is the abbreviation of the reservoir name BERNOULLI Athena, which is composed of the field and formation name. Each sheet has the following seven columns:

- DATE: the last day of the month.
- FIELD: name of the oil field.
- FORMATION: name of the formation.
- WELL_NAME: name of the oil well.
- OIL: monthly oil production rate. Unit: STB/Day.
- WATER: monthly water production rate. Unit: STB/Day.
- GAS: monthly gas production rate. Unit: SCF/Day.

The production data for each well are stacked in rows. For instance, if the first well has 30 months of production history, then the first 30 rows of the data corresponds to the first well. The data for the second well starts from the 31st row, etc.