The extinction of machine learning?

A short note by Dan Cornford on the place of machine learning in the geosciences, part of the *Art of Science* series of technical notes.

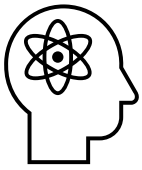
> *Don't get fooled by ML, AI, they are going to fade away in 5 years or less in geology at least.*
>
> Zhiyong He, 23/11/2021, Linked-In

I've recently noticed several discussions on Linked In, and elsewhere, that promote Artificial Intelligence and Machine Learning as the tools to take the oil and gas industry through and beyond the energy transition. This appears, increasingly, to be the view of 'management' in many oil and gas companies – associated with a hope that these tools offer the solutions to (or a distraction from?) many of their problems.

I find this interesting.

My bias (full disclosure)

My training was as a physical scientist in maths and meteorology, I did my PhD in spatial statistics applied to climate and weather, my postdoc in machine learning applied to weather forecasting, and then worked as a Lecturer, then Reader leading research projects in managing uncertainty in complex (physical) systems where I increasingly became convinced that a Bayesian approach to modelling of all forms was the key to clarity of thought and scientific integrity. That took me 20+ years.



I then joined IGI. My aim was to get out of the research rat race, which has become increasingly challenging as universities seek to balance their role as trainers[1] and researchers, while being run as businesses. Much research in universities runs on a 3 year 'PhD' / project cycle and is strongly driven by whatever is 'sexy' at the time, not necessarily what is useful. I wanted to do something more useful, something longer term: to really improve how we use and understand data and models in the natural sciences, especially in the context of oil and gas.

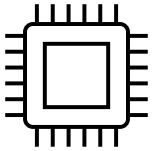
This flurry of interest in machine learning (and AI and data science) got me thinking – is Zhiyong right and this is a flash in the pan in geology, or are the machine learning exponents right, and this will have long term traction, maybe even a central role, in the discipline?

I believe to answer this question one must fundamentally understand the role of models and data in the geosciences.

What are models?

In any real-world study, whether we consider it explicitly, or whether it is implicit in our workflows, we are always building models. I consider a model to be any representation of a system or process which we are studying, abstract or concrete. In a general setting these systems or processes could be anything, for example the purchasing patterns of consumers on a major shopping web site. In the geosciences the systems and processes will typically be real-world, physical systems of which we already have significant (physics-, chemistry- and biology-based) process understanding.

[1] I used to believe we were educators, but for too many students, universities are there to open career paths, not minds



Models do not have to be physics based (such as forward modelling of compressional waves in solids for seismic inversion), they can also be empirical (e.g. the relation between biomarkers in the system and the depositional environment of the source rock) or purely data driven such as the identification of oil families in a basin or region using clustering models. In general, our models will always be supported by some scientific, process-based understanding providing a more robust confidence in any relationships established based on data, especially when extrapolating.

It is very rare that models are context free. We generally build models to answer a question. The question might be around the generation potential of a play, the depositional environmental characteristics of a source rock, the type (and number) of sources of an oil in a reservoir, whether an oil has been water washed, or biodegraded, the migration pathways and mechanisms in a prospect or for CO_2 storage, etc. There are infinitely many questions we can ask, and we will build models to answer these questions.

A key point to always recall, after statistician George Box, is:

All models are wrong, some are useful.

We interpret this to mean all models are approximations of reality. Reality simply is. Models are a human construct, and while the intention here is not to go too deeply into the philosophy of science,

it is good to remind ourselves that **we** are constructing the models. They reflect our beliefs. They are also always approximations. This is true of any model, from a very detailed physics-based simulator of fluid flow in a pipeline (or porous media) to a very simple empirical, statistical model relating bitumen reflectance to vitrinite reflectance.

Using models

It can be helpful to think about modelling in the following (subjective Bayesian) framework:

1. **Research**. You start with some prior knowledge of the problem you are addressing – this can come from a physics / chemistry / biology-based understanding, experience with similar problems in the past, knowledge of the area based on data you have previously seen, and published sources you trust.
2. **Prior formulation**. You formulate a conceptual model that represents your prior knowledge. This could initially be very descriptive, but at some point you will probably formalise this to a computational model – either in a piece of software, or as some code.
3. **Experimental design**. Plan your 'experiment' and identify the data / observations needed to 'learn, or calibrate, your model'.
4. **Learning or inference**. You try and 'learn your model' from new data you acquire (either purchase, go in the field and sample / analyse, or find in the literature / other sources). If you come from a more physics-based background you might talk about 'calibrating' your model, a statistician might talk about 'inference' in their model.
    a. Potentially iterate to 3 if your model answers are not sufficiently informative to allow you to proceed to 5 – this is known as *adaptive experimental design*.
5. **Decision making**. Most real projects result in decisions – models are means to rationally making those decisions, but the decision is the aspect that really matters (unless you are doing pure science, when the model may be the goal).

Since we already accepted all models are wrong, it should come as no surprise here that we will argue data and uncertainty play a key role. We consider learning to be the process of generalising experience, or reducing uncertainty in the outcomes, given a series of 'inputs' (data). In statistics this would be called 'inference', in machine learning, 'learning', in the physical sciences 'calibration' or 'data assimilation'.

When dealing with computational models we characterise learning as:

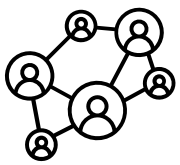
"Reducing our uncertainty in the model representation and outcomes given observational data."

If learning is simply the process of updating our beliefs about our models, what are the differences between machine learning models and physically motivated models? Well, not as much as you might think. For example, a Gaussian process (using a squared exponential kernel), which is a particular class of models popular in machine learning, represents the solution space of functions that can arise from diffusion processes.

Are machine learning and physics-based models different?

Every model can be expressed as **y**=**f**(**x**;**w**)+**e**. Here **y** are the outputs of the model, **x** the inputs (state), **w** the model parameters and **e** the noise or model discrepancy (probabilistic representation of the difference between the model and reality)[2]. The model function, **f**(), could be a linear regression, a complex multi-layered neural network, or the solver for a set of differential operators (representing a physics-based model). In a Bayesian setting the form of **f**() defines our prior beliefs (constraints) over the possible solutions the model should admit.

Simplistically, machine learning models, in all their flavours, are in essence based on some 'smoothness' or continuity assumptions on **f**(), to interpolate between (and extrapolate beyond) observations. There are very few real physical models (above the molecular scale) that are not essentially expressions of conservation equations, with continuity conditions. That combination generates smoothness. It is important to remind ourselves that even well-known physically based models such as the laws of thermodynamics, or Euler equations of flow are actually statistical models – they were originally derived from empirical studies, but are now seen as being based on statistical physics modelling of the interactions of individual atoms or molecules (which do not at the molecular scale obey those equations). At any scale of interest the averaging over 10^23+ molecules means we can treat these equations as excellent fits to observed behaviour. But they are strictly equations for the mean behaviour.



Like religions, for example Catholicism and Protestantism, which are in essence all based on the similar underlying set of beliefs, machine learning and process-based modelling are not so different. And like religions these small differences get amplified to focus on what divides, not what unites. This would appear to be a human condition.

The truth is we can do silly things with any type of model. Each type has its strengths and weaknesses, and the distinction between different types of models is somewhat arbitrary. Understanding that a model defines a prior over **f**(), and that learning and calibration are the same

[2] For notational convenience we assume an additive noise model, however more general formulations are possible, and in some settings necessary.

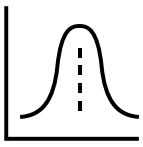
fundamental problem gives us a more powerful way to think about models. It allows us to select the right tool, or **f**(), for the job.

A summary

I would advance the following decision process for selecting whether machine learning or AI methods might be appropriate to solve a problem:

	I have vastly more data points (samples) than variables (parameters) to 'learn'	I have a similar number, or fewer data points (samples) than variables to 'learn'
I know almost nothing about (the response of) the system	Use **machine learning**, e.g. deep neural networks	Use **Bayesian linear methods**, quantify uncertainty
I have some prior knowledge of the system, in terms of the response for my variables of interest	Use **machine learning**, but maybe select model structure more carefully, consider Bayesian methods	Use **Bayesian machine learning** methods and take care to avoid over-fitting
I have prior knowledge of the physical / chemical processes in the system	Understand whether you need to build a full **physics-based** model, or **machine learning** with appropriate models, or **emulation** / **surrogate** models	Build a **physics-based** model. Consider (Bayesian) **calibration** and **data assimilation** to learn about the model using **Bayesian emulation**
I have prior knowledge of the physical / chemical processes in the system but I am only interested in some small number of inputs and outputs	Consider an **emulation** or surrogate modelling approach combine physics-based and statistical models	Consider **Bayesian emulation** / surrogate modelling approach, thinking carefully about uncertainty quantification

The above table is wrong. Hopefully it is still useful. I'd argue in most natural science disciplines we are typically in the right-hand column – data sparse. A reasonable question would be what does "vastly more" mean. There is not a simple answer here – it depends on many factors, such as the uncertainty in the observations, the complexity and dimensionality of the model, but as a rough rule of thumb for a non-linear model I'd want roughly 10 (independent) observations for each parameter I am trying to estimate, ideally covering the region of interest.



For me the key point is that the question should not be machine learning or process-based modelling, rather how do we best combine these to facilitate understanding and decision making. A deep understanding is necessary to see models, whether they be 'statistical' or 'physically based', as in essence being the same thing. It is in fact more important to acknowledge that a proper treatment of uncertainty is the most important consideration in our subsurface context, not the set of tools we use to help us generalise from our data. The representation and treatment of uncertainty will form the basis of a future piece.

I believe machine learning, especialy in a statistical framework will have a long and fruitful role in our toolkit alonside a range of other models, whether they be physically motivated or not. Machine learning, like all scientific developments is subject to a hype-cycle, and is probably somewhere near the 'peak of over-inflated expectations' in most areas of geoscience. But it will reach the 'plateau of productivity', and arguably is doing so in some data-rich areas, such as seismic processing.

To update Box's quote:

All Linked-In posts are opinion, some are useful.

I leave you to be the judge of that.