Machine learning in geoscience notes

January 23, 2025

Chapter 1

Introduction

In our relentless quest for knowledge and to improve our lives, humanity has harnessed an elaborate enterprise of knowledge, we call science. Over the centuries, luminaries like Al Khwarizmi, Newton, Fermat, and Pascal have contributed to its evolution, shaping an expansive legacy, now embodied in countless books, journals, universities, research organizations, and more. Science, over the years, has paved the way for immense developments to improve our lives. Now, science has brought about a transformational development in the form of machine learning, poised to challenge and refine much of our traditional theory-based understanding. At its core, machine learning advocates for extracting knowledge directly from data, shifting the paradigm away from purely theoretical constructs. This shift extends to our understanding and predictions of the Earth's contents and dynamics. Will science's future be guided by data, or will theoretical advancements remain essential? In this course, we will explore the fundamentals of machine learning and strive to strike a meaningful balance between the power of data and the enduring value of geoscientific knowledge. At the heart of this geoscientific knowledge is the Earth's Symphony.

1.0.1 The secrets of the Earth's Symphony

Beneath the surface of the Earth lies a vast, uncharted world, a symphony of shifting plates, whispering sediments, and resonating echoes. The Earth, with its endless dynamism, has long guarded its secrets. Yet, in the depths of its dark chambers lie the keys to locating natural resources, mitigating disasters, and revealing the stories of epochs past. Today, armed with data of measured echoes, as well as electrical and magnetic responses, and our expectations of these hidden chambers based on our experiences and preconceptions, we are unveiling the Earth's concealed truths, and illuminating its dark chambers. Equipped with the physics behind the measurements and harnessing the power of inverse theory, geoscientists are unlocking these mysteries, deciphering seismic signals, and other physical measurements, with relatively high precision that was once unattainable. Cumulating into an otherwise unimaginable hypothesis that, for

one, the Earth's outer core, 2900 km deep, is fluid. All of which are garnered under the umbrella of science.

However, science is incomplete, and data/measurements are always imperfect, leaving us with wide gaps of knowledge in unveiling the Earth's concealed truths. These gaps have limited our ability to predict Earthquakes, discover additional resources, and fully understand the Earth, among other desires. The uncertainty involved in characterizing the Earth's content is large. Though we can predict that the Earth inner core is most likely fluid, as an example, we are not sure of the mantle -outer core boundary depth, and we are certainly not sure of the locations of plumes in the mantle. The quantification of uncertainty in our predictions is as important as the predictions themselves. In some cases, even more important. Predictions and the statistics behind them are at the core of the machine learning paradigm, and thus we hope to be able to garner the power of machine learning to make better predictions and assess our confidence in them. This objective is meant to strike the balance between science and data (Figure 1.1), and specifically to garner machine learning as a numerical tool to solve our outstanding geoscience problems.

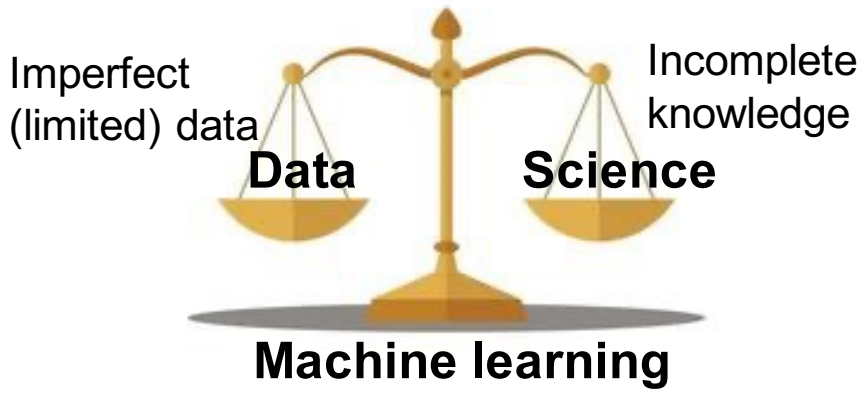

Figure 1.1: The balance between data and science.

1.1 Learning machines or machines that can learn

The quest for enabling devices to perform various tasks can be traced back to the dawn of science. In the evolution process of our existence on this planet, intelligence has always been synonymous with humans. This ability to acquire, process, and apply knowledge and skills to solve problems, adapt to new situations, and learn from experience. However, our ability to do so depended on our intelligence, which is finite at best and reasonably limited. So, through

using our collective, accumulated intelligence as humans, we decided to pass the intelligence baton onto machines that we hope will do a better job of it and assist us in solving outstanding problems, as well as making our lives easier. Thus, we have the term Artificial Intelligence (AI). It is considered a branch of computer science focused on creating systems or machines capable of performing tasks that typically require human intelligence. AI often leverages algorithms, data, and computational power to perform these tasks that can include problem solving, learning, reasoning, perception, language understanding, and decision making.

Within the AI universe, machine learning (ML) has gained wide praise in recent years for its practical uses and successes. The concept behind ML involves an important component of AI, which is learning from experience. In particular, **it focuses on creating algorithms and models that enable computers to learn patterns and make decisions or predictions from data without being explicitly programmed**. So, most ML frameworks involve training, where the machine gains experience over time, and then we use this machine-acquired experience to make predictions. At the heart of the training process are the data (examples) from which the machine learns. Machine learning is widely used in applications such as image recognition, natural language processing, recommendation systems, and autonomous vehicles.

In a mathematical sense, ML lies at the intersection of statistics, calculus, linear algebra, optimization, and numerical methods. The training relies heavily on inverse theory. The neural network model that is being trained extracts its algorithmic form mainly from linear algebra. The training process relies on numerical methods. The iterative optimization process involves the calculus of gradients and the chain rule. Finally, the performance of the model depends on the statistics of the training data. So, ML involves a rainbow of mathematical knowledge, yet its main objective is prediction, and in this course, predictions in geoscientific applications.

1.2 The Intersection of Geoscience and machine learning

ML is not just a tool, but a paradigm shift in the way we approach tasks in natural language (NL), in image generation, in speech analysis, and everything in between. Geoscience is no exception. Geoscience, in particular, can benefit enormously from the ability of ML to handle large datasets, recognize patterns, and make predictions. From identifying subsurface structures to predicting earthquake occurrences, ML serves as both a magnifying glass, as well as a compass, of our data.

To appreciate the synergy between geoscience and ML, consider the fundamental law governing seismic wave propagation: the wave equation. At the heart of many geophysical processes lies the acoustic wave equation, which specifically

describes how seismic waves propagate through acoustic media:

$$\nabla^2 u(\mathbf{x}, t) - \frac{1}{v^2(\mathbf{x})} \frac{\partial^2 u(\mathbf{x}, t)}{\partial t^2} = s(\mathbf{x}, t), \tag{1.1}$$

where $u(\mathbf{x}, t)$ represents the wavefield (displacement or pressure), $v(\mathbf{x})$ is the velocity of wave propagation, dependent on the medium, $s(x, t)$ is the source term, describing the seismic source, $\mathbf{x}$ is the space coordinates, t is time, and ∇^2 stands for the Laplacian operation.

Traditional methods solve this equation using numerical approaches, such as finite-difference or finite-element methods. Although effective, these techniques often require significant computational resources for accurate and stable approximate solutions. ML, however, offers alternative methodologies, in which models learn to approximate solutions or infer properties directly from data, bypassing some of the limitations of conventional approaches.

The relation between science (geoscience) and machine learning over the years can be encapsulated in the following historical perspective. Neural networks are predictive models, and one of their roles is to use data to predict what happens next from patterns they learned in the data. Despite that machine learning is relatively new (effectively defined in the 1950's), interestingly, scientists used data, going back to the 1600s, in a similar predictive fashion. For example, Johannes Kapler and Galileo Galilei used these to monitor and gather data on planet motion and used them to predict the planet's trajectory. Soon after, physical laws that describe motion with respect to gravity, thanks to scientists like Isaac Newton, mitigated the need for such prediction practices. We can now estimate the trajectory of planets by solving simple equations. This type of transition from predictive models to physical laws has been a trend in the past 400 years (depicted in Figure 1.2), and physical laws became king. Suddenly, in the past two decades, we seem to be back to predictive models thanks to machine learning. What happened and why did it happen? Are we abandoning physical laws? These questions are at the heart of the new transition.

1.3 Seismic Data: A Treasure Trove

Data, specifically our observed and measured, are the cornerstone of scientific inquiry and the accumulation of knowledge, serving as the foundation for hypothesis testing, theory development, and evidence-based decision-making. In essence, measured data are not just a tool for science; it is the essence of how science progresses. Geophysical and seismic data are no exception. As a result, we have accumulated data for years. In fact, at one point, seismic data were only second to meteorological data in the amount acquired. So, we always had data collected, usually from observations and measurements. The role of ML revolutionized how we deal with data (see Figure 1.3).

Seismic data, typically recorded as time-series signals, are the primary input for geoscientific investigations. These signals contain rich information about the

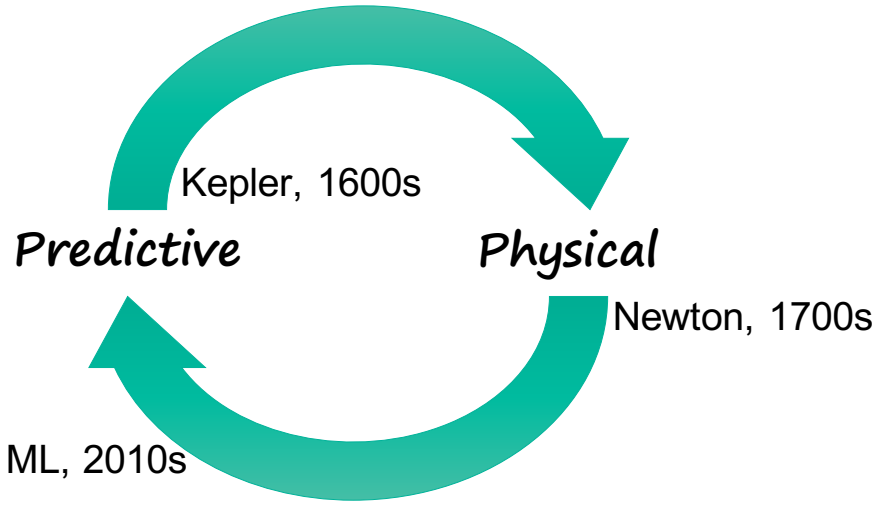

Figure 1.2: The cycle of predictive versus physical models over the past 5 centuries.

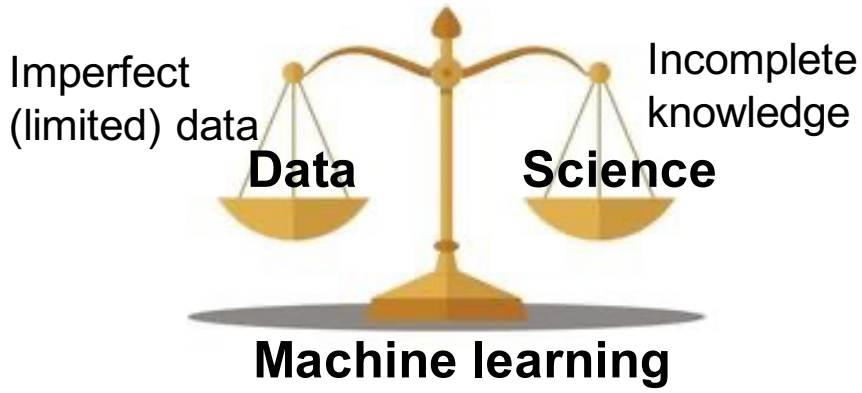

Figure 1.3: How data processing has evolved over the decades. We started with using our knowledge and experience to manually analyze data we visualize and then perform predictions, then we started to use our knowledge to build fixed algorithms that perform such analysis faster, and predict. Finally, with ML, we use algorithms that can gain the experience itself and do the predictions. Maybe no knowledge required!

subsurface but are often obscured by noise and complexities. The challenge lies in extracting meaningful features. This is where ML excels.

Consider a seismic trace, $s(t)$, which can be modeled as:

$$s(t) = \sum_{i=1}^{N} a_i h(t - \tau_i) + n(t), \tag{1.2}$$

where a_i are amplitudes associated with reflections, $h(t)$ is the source wavelet, τ_i are the travel times of reflected waves, and $n(t)$ represents noise.

ML algorithms, particularly deep learning models, can disentangle these components, separating signal from noise and identifying patterns that correspond to subsurface features, as well as enhance the resolution of reflections.

1.3.1 Applications and Opportunities

Machine learning has revolutionized several key areas of seismic analysis. Here is a list of some examples in that regard:

1. Seismic Inversion: ML models approximate subsurface properties, such as density and velocity, by learning mappings from seismic data to physical parameters.

2. Fault Detection: Convolutional neural networks (CNNs) excel at identifying fault lines within seismic images, outperforming traditional edge-detection methods.

3. Earthquake Prediction: Recurrent neural networks (RNNs) and transformers analyze time-series data to forecast seismic events with improved accuracy.

4. Seismic Imaging: Generative models, such as variational autoencoders (VAEs) or GANs, reconstruct high-resolution images of the subsurface from sparse or noisy data.

1.4 The ML paradigm as an optimization problem

At the heart of the ML framework is an optimization problem; to specifically find the optimal ML machinery (parameters or weights) that can make it accomplish the desired task. Thus, at its core, machine learning (ML) can be understood as the process of optimizing a model to perform a specific task based on data. This optimization-centric perspective is particularly powerful for tackling complex, multivariate, and often noisy datasets encountered in geosciences.

"The future influences the present just as much as the past" Friedrich Nietzsche. In other words, our expectations of what will happen in the future influence our current decisions as much as our experiences of the past. This also holds for our predictions in geoscientific applications. So, the optimization nature of ML allows us to incorporate these components into our predictions, as

well as other components. So, fundamentally, ML gains experience from training of data, thus defining the term data driven. However, we can incorporate our expectations and other sources of information, like well-information. In fact, we can incorporate our scientific knowledge, and specifically our physical laws, into the process. Figure 1.4 harnesses the agility of the optimization framework in including these components in the prediction, and even taking into account the inductive bias of machine learning models.

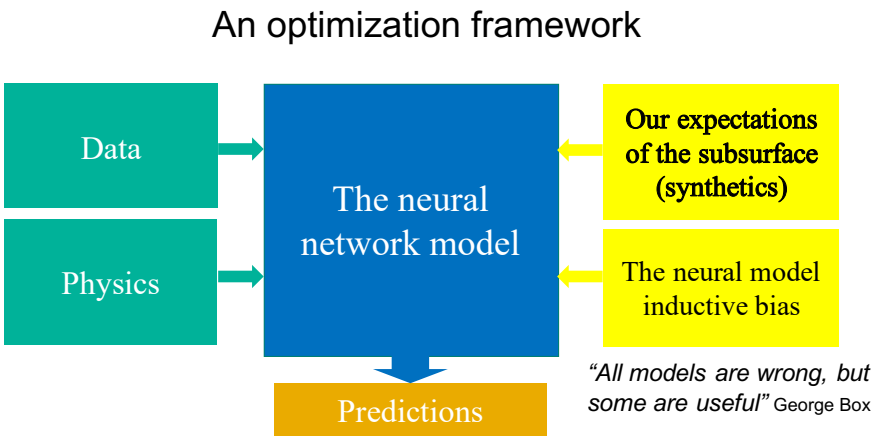

Figure 1.4: The optimization problem nature of ML and what can we incorporate into it from geoscience.

Understanding machine learning as an optimization problem provides a unifying framework that bridges theory and practice. It enables geoscientists to design models that not only achieve high performance but are also interpretable and aligned with physical laws. As we delve deeper into applications, this perspective will serve as a foundational lens through which complex geoscientific problems can be approached and solved.

1.5 The short story of how ML evolved

The first evidence of machines performing tasks can be traced back to the work of Blaise Pascal, a French mathematician, physicist, and inventor, who built an early mechanical device designed to perform mathematical tasks [Dawson(1970)]. Specifically, in 1642, Pascal invented the Pascaline, one of the first mechanical calculators. Pascaline was designed to perform additions and subtractions, and it could also be used for multiplication and division through repeated additions or subtractions.

The device used a series of gears and dials to represent numbers, and it was a significant step forward in the development of computing machines. Pascal's invention was primarily motivated by the need to assist his father, who was a

tax collector, in performing calculations more efficiently. Although the Pascaline was not widely adopted because of its complexity and limitations, it laid the groundwork for future developments in mechanical computing and is considered a precursor to modern calculators and computing devices.

However, actual ML roots trace back to the mid-20th century [Hauser(2009), Goodfellow et al.(2016)Goodfellow, Bengio, and Courville], beginning with Alan Turing's 1950 concept of a "learning machine" in his seminal paper "Computing Machinery and Intelligence". In 1959, Arthur Samuel coined the term "machine learning" while developing a program to play checkers that improved over time. The 1960s and 1970s saw foundational developments in neural networks, including Frank Rosenblatt's invention of the perceptron in 1958, which was an early algorithm for supervised learning. However, progress stalled due to computational limitations and criticisms, such as those in the book "Perceptrons" (1969) by Minsky and Papert, which highlighted the limitations of single-layer neural networks. In the 80s, the interest accelerated, spurred by advances in algorithms like backpropagation, which allowed multi-layer neural networks to learn effectively. This era also saw the rise of probabilistic methods, such as Bayesian networks, laying the groundwork for modern ML approaches. In the 90s, the focus shifted to data-driven methods, with the explosion of digital data enabling algorithms like support vector machines (SVMs) and decision trees to thrive. This period also marked the emergence of ensemble methods, such as boosting and bagging, which improved predictive accuracy.

The 2000s and 2010s saw a dramatic leap with the advent of deep learning, powered by increased computational power, large-scale datasets, and advanced hardware like GPUs. Breakthroughs in convolutional neural networks (CNNs) for image recognition and recurrent neural networks (RNNs) for sequential data revolutionized fields like computer vision, natural language processing, and speech recognition. Today, machine learning continues to evolve, with research focusing on interpretability, ethical AI, reinforcement learning, and generative models such as transformers and diffusion based networks. ML has become a cornerstone of technology, transforming industries and shaping the future of AI-driven innovation.

However, the evolution of ML was not always a smooth ride. It has experienced several periods of stagnation, often referred to as "AI winters" characterized by reduced funding, research activity, and enthusiasm. These winters occurred when the technology failed to meet inflated expectations, leading to disappointment among researchers, funders, and the public.

The first winter occurred between 1974 and 1980. Early AI systems, such as rule-based expert systems and the perceptron, showed promise but failed to deliver practical solutions due to computational and theoretical limitations. The release of the Lighthill Report in 1973 criticized the field's progress, prompting governments, particularly in the U.S. and the U.K., to reduce funding for AI research.

The second AI winter spanned from 1987 to 1993, during which expert systems, once heralded as revolutionary, were exposed as brittle and costly, requiring extensive manual input to function and lacking adaptability. The failure

of Japan's ambitious "fifth generation" computing project further diminished confidence in AI's potential. As a result, investments and research momentum in the field decreased sharply.

The early 2000s slowdown was less severe but marked a lull in the adoption of neural networks. Techniques like support vector machines (SVMs) and decision trees were more practical and efficient at the time, leading to a temporary stagnation in neural network research. This period persisted until the resurgence of interest in deep learning in the mid-2000s, driven by advances in computational power, data availability, and algorithmic improvements.

These winters highlight the cyclical nature of AI development, where periods of optimism and investment are often followed by setbacks when expectations outpace technological capabilities.

1.6 ML history in geoscience

Among the first work to utilize ML in a geoscience applications can be traced back to the early 90s [McCormack(1991)]. Soon after [Röth and Tarantola(1994)]

The application of machine learning (ML) in geoscience has evolved significantly over the past few decades [Bergen et al.(2019)Bergen, Johnson, de Hoop, and Beroza, Luan and Tian(2022)]. In the early stages, statistical methods such as linear regression and principal component analysis (PCA) were used for tasks like geophysical inversion and reservoir characterization. By the 1990s, advances in computing power enabled the adoption of more complex algorithms, including neural networks, which were applied to seismic interpretation and prediction of petrophysical property [McCormack(1991)]. It also gained an attraction in seismic inversion applications [Röth and Tarantola(1994)].

In the 2000s, support vector machines, decision trees, and ensemble methods gained popularity for geospatial data classification and environmental modeling. The 2010s marked a transformative era with the rise of deep learning, driven by increased computational resources and the availability of large datasets. Techniques like convolutional neural networks (CNNs) revolutionized seismic data analysis (including denoising), while recurrent neural networks (RNNs) were used for time-series prediction in earthquake forecasting and climate modeling, as well as FWI gradient conditioning [Sun and Alkhalifah(2020)].

Today, ML continues to transform geoscience, enabling breakthroughs in subsurface imaging, fault detection, lithology classification, and resource exploration. It has also spurred the development of hybrid models that integrate physical knowledge with data-driven approaches, fostering innovation across geophysical research and industry applications.

1.7 Challenges and Ethical Considerations

Although ML promises remarkable advances, keep in mind that it is not without challenges. Overfitting, data bias, and the "black-box" nature of many

algorithms pose significant hurdles. In addition, ethical considerations, such as the use of ML for resource exploitation, must be addressed responsibly.

Let us take these challenges one point at a time. Overfitting is a popular term in ML that implies that our trained model knows how to work well on the known training data but not as well on new data, which is the purpose of our training. We will cover this phenomenon in the course in detail. Overfitting can cause the network to produce predictions that look realistic, but inaccurate. Specifically, it leads to poor generalization, where the model performs well on the training data but fails to make accurate predictions on unseen data. This results in overly confident or incorrect predictions during inference, as the model captures noise or irrelevant patterns from the training data. Additionally, overfitted models are more sensitive to slight variations in input, reducing their robustness and reliability in real-world applications.

Likewise, data bias involves the training aspect, and it implies that we have more data samples of one type (like one class or two) than other types. This will bias the model toward features that are defined by the class or the corresponding feature that is better sampled. For example, in image classification, data bias in machine learning training can lead to systematic errors during inference, as the model may produce skewed or unfair predictions that reflect the bias present in the training data. This can result in discrimination against certain groups or inaccurate outcomes for underrepresented or overrepresented classes. Furthermore, biased models can perpetuate or amplify societal inequalities and may not be able to generalize effectively across diverse real-world scenarios.

On other hand, the "black-box" nature of machine learning models makes it difficult to understand how predictions are made during inference, which can erode trust and confidence in their decisions. This lack of transparency complicates the identification of errors or biases in predictions, making it challenging to improve or validate the model's reliability. Moreover, it hinders interpretability and accountability, especially in critical domains like healthcare or finance, where explanations are crucial for ethical and informed decision making.

Machine learning can amplify biases in training data, leading to unbalanced (or unfair) and discriminatory outcomes that disproportionately impact marginalized groups, or in our case rarely occurring features. The lack of transparency in model prediction process raises concerns about accountability and trust in the results, especially in high-stakes applications such as hiring, criminal justice, and healthcare. Additionally, the use of personal data in training can infringe on privacy rights and lead to misuse or unintended exploitation of sensitive information.

1.8 Course objectives

This course aims to equip students with the theoretical foundations and practical skills necessary to apply machine learning (ML) techniques to geoscientific challenges. By the end of the course, students will understand the fundamental principles of supervised, unsupervised, and deep learning algorithms and their

relevance to geophysical data analysis and modeling. They will gain proficiency in handling geoscience-specific datasets, such as seismic data and learn how to preprocess and analyze these datasets effectively.

Students will explore real-world applications of ML in geosciences, including tomography, denoising, and seismic inversion. The course emphasizes the integration of domain knowledge with machine learning to build interpretable and robust models. Through hands-on projects, participants will develop the ability to implement ML workflows using programming tools and libraries (primarily Pytorch), evaluate model performance, and address challenges such as overfitting, data bias, and uncertainty quantification.

Ultimately, this course aims to prepare students to innovate and contribute to the rapidly evolving intersection of machine learning and geoscience, enabling them to solve complex problems in academia, industry, and beyond.

1.9 A Vision for the Course

This course explores the intersection of machine learning and geoscience, with a focus on seismic applications. Through theoretical foundations, practical examples, and case studies, you will gain the skills to:

- Understand the main concepts involved in machine learning. - Develop and apply ML models to geoscientific problems. - Critically evaluate ML solutions within the context of real-world challenges.

As we journey through these topics, remember that the symphony of the Earth is vast and complex, and each discovery is but a note in the grand composition. With ML as a tool, we are on the brink of profound understanding.

Bibliography

[Bergen et al.(2019)Bergen, Johnson, de Hoop, and Beroza] K. J. Bergen, P. A. Johnson, M. V. de Hoop, and G. C. Beroza. Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363(6433):eaau0323, 2019. doi: 10.1126/science.aau0323. URL https://www.science.org/doi/abs/10.1126/science.aau0323.

[Dawson(1970)] R. E. Dawson. *Blaise Pascal: Mathematician, Physicist, and Thinker about God*. Enslow Publishers, Inc., 1970. Discusses Pascal's contributions to science, including his invention of the Pascaline.

[Goodfellow et al.(2016)Goodfellow, Bengio, and Courville] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[Hauser(2009)] S. Hauser. The history of artificial intelligence. *AI Magazine*, 26(4):53–68, 2009.

[Luan and Tian(2022)] Q. Luan and Z. Tian. Application of machine leaning methods in geoscience. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 563–567, 2022. doi: 10.1109/ICETCI55101.2022.9832203.

[McCormack(1991)] M. D. McCormack. Neural computing in geophysics. *The Leading Edge*, 10(1):11–15, 1991. doi: 10.1190/1.1436771. URL https://doi.org/10.1190/1.1436771.

[Röth and Tarantola(1994)] G. Röth and A. Tarantola. Neural networks and inversion of seismic data. *Journal of Geophysical Research: Solid Earth*, 99(B4):6753–6768, 1994. doi: https://doi.org/10.1029/93JB01563. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93JB01563.

[Sun and Alkhalifah(2020)] B. Sun and T. Alkhalifah. Ml-descent: An optimization algorithm for full-waveform inversion using machine learning. *GEOPHYSICS*, 85(6):R477–R492, 2020. doi: 10.1190/geo2019-0641.1. URL https://doi.org/10.1190/geo2019-0641.1.