

ADVANCED DATA ANALYTICS IN GEOTECHNICS

Adapting to the Big Data Era

By Nick Machairas, PhD, A.M.ASCE, and Magued Iskander, PhD, PE, F.ASCE



Everyone is talking about the impact of artificial intelligence (AI) on the future of work. In one camp are people like the industrialist Elon Musk who fear the potential evils of AI and its possible impact on the future of mankind as depicted in the movies *The Terminator* and *The Matrix*. In the opposite camp, people like Nobel laureate Joseph Stiglitz claim that AI will usher a new age of prosperity that rivals the industrial age. The truth is, of course, somewhere in between. In this article, we provide a primer on AI, debunk some misconceptions, and explore some of the possible impacts of AI on the future of geotechnics.

In today's rapidly evolving technological landscape where innovations can become irrelevant within days, advances in AI are constantly making headlines, sparking awe, controversy, and sometimes both. In reality, though, AI has increasingly gained a foothold in many human spheres, from Siri/Alexa digital assistants to mapping apps such as Waze and shopping apps that offer suggestions based on our purchasing histories. Undoubtedly, technologists, aided by state-of-the-art data analytics, have transformed many industries, producing remarkable predictions and insights. Of interest to us, however, is the question, how can the geoprofession adopt advanced analytics? How can inquisitive geoprofessionals augment their skills and deliver powerful insights and predictions, thus delivering value for their clients and organizations?

What Is Advanced Data Analytics?

Advanced analytics involves the hands-off (automatic), or limited-interaction (semiautomatic), processing of data with the goal of producing insights, predictions, or recommendations. The process can be broken down into two parts: a) efficient automation (e.g., designing custom algorithms for data pre/post-processing, data/text mining, deterministic analyses, visualization, simulations), and b) artificial intelligence (AI) (e.g., probabilistic analyses, machine learning (ML), deep learning). An analytics project may produce remarkable insights without AI; therefore, part "b" is not always included in advanced analytics.

Fundamentally, intelligence is the ability to acquire and apply knowledge and skills. General intelligence is human-level intelligence and cognitive reasoning (i.e., how information is used to make a decision or reach a conclusion). AI is intelligence demonstrated by machines. It's often described by introducing the concept of *intelligent agents*, devices that have a goal and are able to perceive their environment and react to inputs in ways that maximize their chances of success.

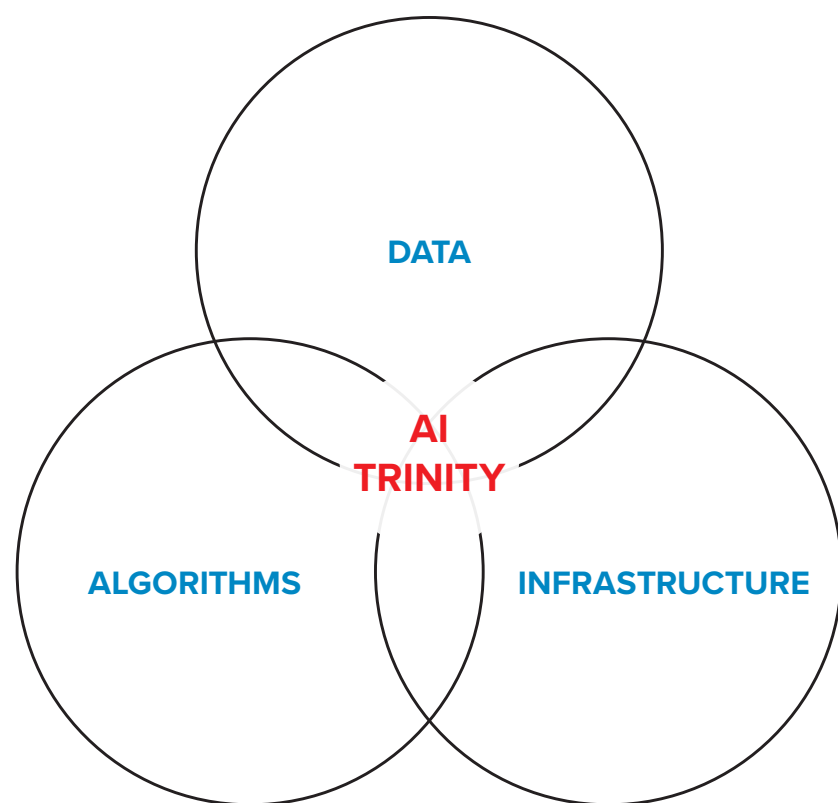


Figure 1. The AI Trinity (adopted from Anandkumar).

AI is intelligence demonstrated by machines. It's often described by introducing the concept of *intelligent agents*, devices that have a goal and are able to perceive their environment and react to inputs in ways that maximize their chances of success.

are able to perceive their environment and react to inputs in ways that maximize their chances of success. For example, an autonomous vehicle has the goal of driving safely from point A to point B, and it's expected to react to unknown conditions in order to do so, without being explicitly programmed. The question that normally arises is how did the vehicle "learn" to safely respond to new inputs. Extending the example from vehicles to intelligent computers processing geotechnical data, the question becomes how to use AI to make useful predictions and/or recommendations.

It's important to first understand how AI works and what factors have contributed to its recent remarkable advances, despite the fact that AI has been around for decades. Professor Anima Anandkumar from Caltech first introduced the concept of *The AI Trinity: Data + Algorithms + Infrastructure* (Figure 1). The proliferation of open datasets and the exponential growth of data are the fuel to machine intelligence. Next, design and implementation of AI algorithms have expanded from previously limited and very expensive operations that became that way because hard-to-find experts and specialized tools were required. Now, however, the barrier to entry has been dramatically lowered with major tech companies and academic institutions open-sourcing their powerful AI frameworks (e.g., Google Tensorflow). Thus, world-class research and sophisticated product development is, in principle, possible by anyone. Finally, training of AI algorithms usually has significant computational requirements that cannot be met by consumer desktop computers. Rather, AI algorithms depend on high-performance computing facilities for training. In the past, these facilities were the exclusive realm of well-capitalized large enterprises and academic institutions, but today, cloud computing has made advanced system architectures and virtually unlimited

computational resources available to anyone. It's therefore clear that the *AI Trinity* is no longer limited to a privileged few; instead, it's available to individuals and companies with varying budgets and objectives.

What's the Status of AI Adoption in the Geoprofession?

The principles of AI were established decades ago. Data, algorithms, and large-scale computing infrastructure are now more readily available than ever. But have geoenvironmental researchers and professionals been using AI? And if so, what's the level of adoption compared to other fields?

While there's clearly a lot of interest, as evidenced by the increasing number of lectures, conference presentations, and workshops in the last few years, it's difficult to reliably quantify the use of AI in the geoprofession. There are some shining examples: the Crystal Ball Workshop held in advance of the fall 2019 meeting of the Geoprofessional Business Association (see "Our GeoBusiness Future - Big Data, Machine Learning, and AI," pp. 40-45), and the International Symposium in

Offshore Geotechnics (ISFOG 2020), scheduled for August 2020 at the time of this writing. Plans for ISFOG 2020 include a pile driving prediction event, hosted on Kaggle, a platform for predictive modelling and analytics competitions. Kaggle permits posting data and user competitions to produce models for predicting and describing the data. These first-of-their-kind events identify difficulties and opportunities for professionals while demonstrating that geotechnical and foundation data can be made readily available for events similar to widely popular competitions in data science. They're noteworthy because they revolve around how to improve business productivity and/or profitability through data and AI rather than their normal academic framework.

To gauge the adoption of AI in geotechnical engineering, we entered the keyword phrases "artificial intelligence" or "machine learning" dating back to 1960 into Scopus, a large abstract and citation database of peer-reviewed literature. The results were aggregated by decade and grouped by several engineering and scientific disciplines, such as computer science, mathematics,

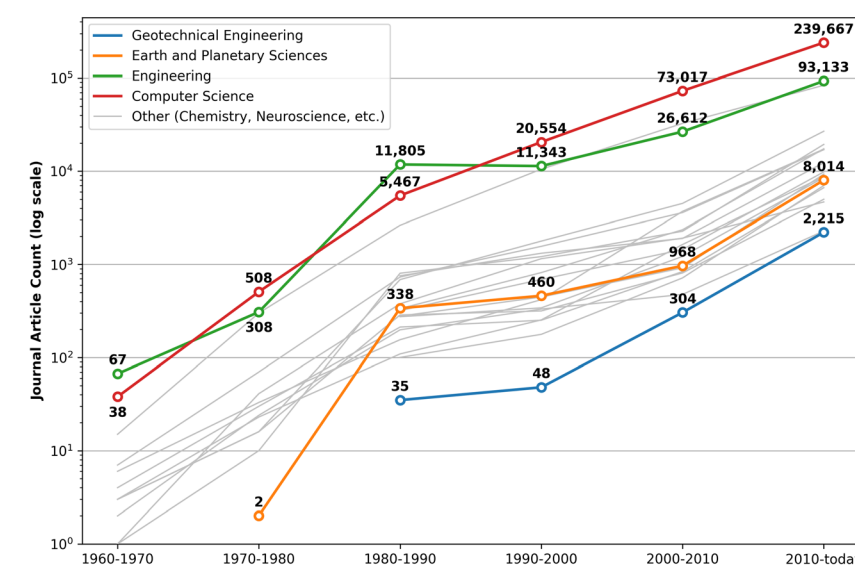


Figure 2. The timeline of research in geotechnical engineering and other fields using AI or ML.

social sciences, medicine, geotechnical engineering, and more (Figure 2). For comparison, geotechnical engineering, Earth and planetary sciences, and the broader engineering field are highlighted in the figure. Article count alone might not necessarily be a sufficient qualitative measure; however, the much lower count for geotechnical engineering is a clear indication of lack of research activity on methods involving AI or ML. While these results might seem discouraging, they highlight the great potential that exists for disrupting existing processes, especially those that generate lots of data, such as site investigations, instrumentations, and performance monitoring.

Industry-adopted innovations often originate from academia, so it would be reasonable to assume that adoption of AI in the geoprofession follows a similar trend. Research, however, is often driven by the priorities of major funding agencies, so the dearth of AI publications in geotechnical engineering is likely related to the lack of emphasis on AI by traditional funding agencies.

What Skills Are Required?

Having witnessed the exponential growth of Big Data, extremely large data sets that can be analyzed computationally to reveal patterns and trends, it's becoming increasingly difficult to process large datasets with traditional software such as Microsoft Excel. This difficulty may be compounded by the near impossibility of setting up custom analyses and simulations with commercially available tools. Today, data professionals rely on powerful open-source tools and must be fluent in one or more computer programming languages to effectively conduct analyses on large volumes of complex data. At the moment, Python is the most popular language for data analysis.

The first step for geoprofessionals interested in using data analytics in their practice is to learn the basics of a programming language, preferably

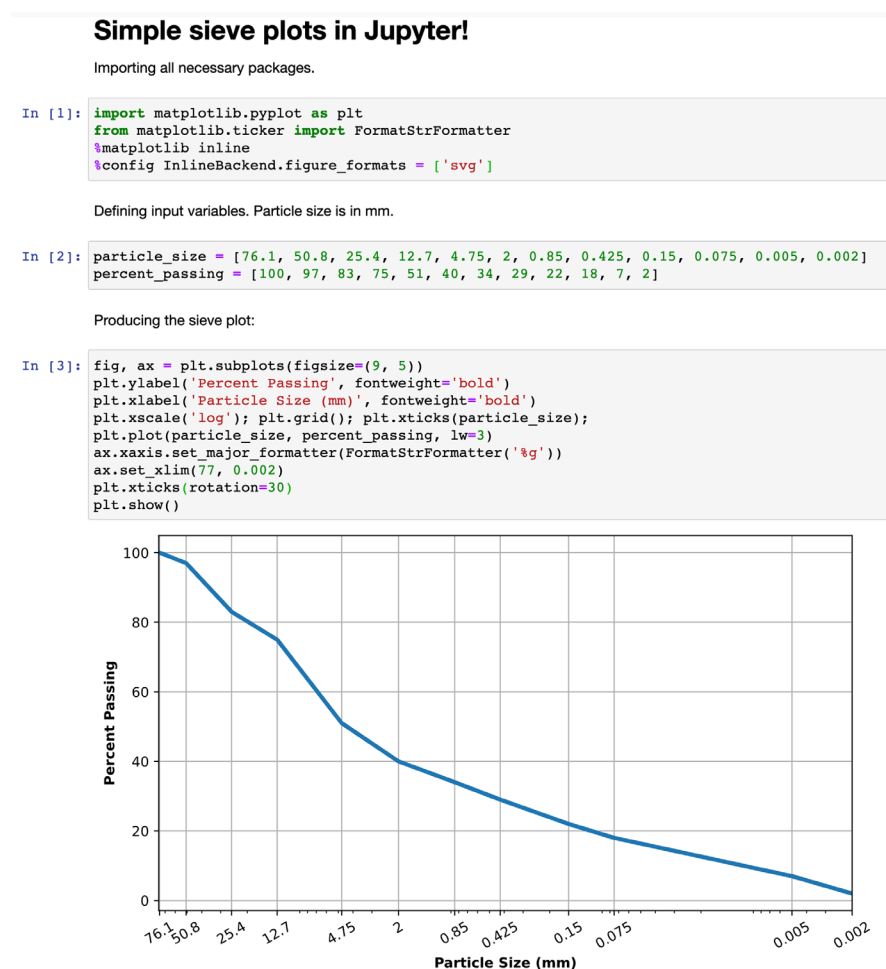


Figure 3. An example of a Jupyter notebook with Python code plotting sieve analysis results.

Python. This may seem a daunting task, but fear not. When it comes to purpose and delivery, a programming language is no different than a spoken language. In fact, the case could be made that learning a new computer programming language is easier for geoprofessionals than learning a foreign language because engineering design is fundamentally algorithmic, and engineering professionals have years of experience in algorithmic thinking through education and experience.

While proficiency in Python is recommended, it's not required, and there's a big difference between *scripting* and *programming*, with the former being easier and less demanding for

beginners than the latter. With Python scripting, users define the computer commands needed to perform a task from start to finish following a turnkey approach, often without paying much attention to reproducibility, expandability, efficiency, or code testing. As an analogy, consider this small data-processing task: a user has several lab results at hand and wants to produce plots for each project to help visualize site conditions. Using a common spreadsheet application, the user interacts with its graphical interface, clicking on the buttons to load the lab data and produce the desired plots. The process is then repeated for all available data files. The user can be far more

efficient and error-free with this task by producing minimal Python code that automates the process. Essentially the “clicks” are replaced by simple computer commands. This might not make much sense when handling a few data files; however, it's not uncommon for large construction projects to produce hundreds of test results that need to be processed and analyzed with a short turn-around time.

Code scripting is used in the vast majority of analytics projects. Computer code may be written in plain text files and then executed, or written and executed piece-by-piece within interactive environments such as the widely popular *Project Jupyter*. This open-source initiative produced *Jupyter Notebook*, a web application that allows users to create and share documents that contain live code, equations, visualizations, and text with each element contained within a “cell.” A simple example of a Jupyter notebook using Python code to plot the results from a sieve analysis is shown in Figure 3. Programming novices will notice that the code is easily comprehensible. And while this is an example of a single plot only, some minor adjustment to the code can make it produce multiple plots with minimal effort.

An especially powerful data asset management process can be advanced

by combining code scripting with the Data Interchange for Geotechnical and Geoenvironmental Specialists (DIGGS). DIGGS (see “What Does DIGGS Do for Me? Better, Faster, Cheaper Is the Goal,” pp. 54-59) is a standard that's designed to help geoprofessionals store and transfer geotechnical data in a manner that promotes data consistency and collaboration. Data are stored in XML format, making it ideal to process with many programming languages, including Python.

Getting Started

How can you get started with AI? Online learning has revolutionized skill building. Most academic institutions offer a number of high-quality course options online, either in-house or through non- or for-profit online education platforms (*edX* and *Coursera* are popular). A motivated individual can evolve from absolute beginner to advanced programmer in a matter of weeks to months. Credentials earned from quality online platforms can be excellent resume boosters. The authors list a number of useful resources on their website, wp.nyu.edu/iskander/resources.

AI

The skills we've discussed so far relate to the “Efficient Automation” part of Advanced Data Analytics. This skill set alone can be very useful for a multitude

of geoengineering tasks that go beyond simple plots (i.e., data mining, preprocessing, cleaning and running complex analytical procedures on very large datasets). But learning how to use AI for predictive analytics can be trickier. First, AI is a popular news item, so it's often used by online tutors as a way to gather attention to their business. There are countless videos and blogs that oversimplify the use of AI. This is good for lowering the level of entry, but they may lead to a false sense of expertise. While there are excellent guides on the internet, an AI novice should always exercise critical thinking, adopting reliably good techniques but rejecting those without proper justification. Probabilistic methodologies require, as the name implies, a strong foundation on probabilities and inference. Hence, users should dust off their foundational knowledge in these areas before graduating to AI. There are many excellent, and free, online courses that can help with self-paced studying. There are also comprehensive courses that cover probabilities in the first weeks of their syllabus before moving on to ML algorithms.

ML

ML is a subset of AI. ML algorithms make predictions for the future by learning from past experiences without relying on preprogrammed

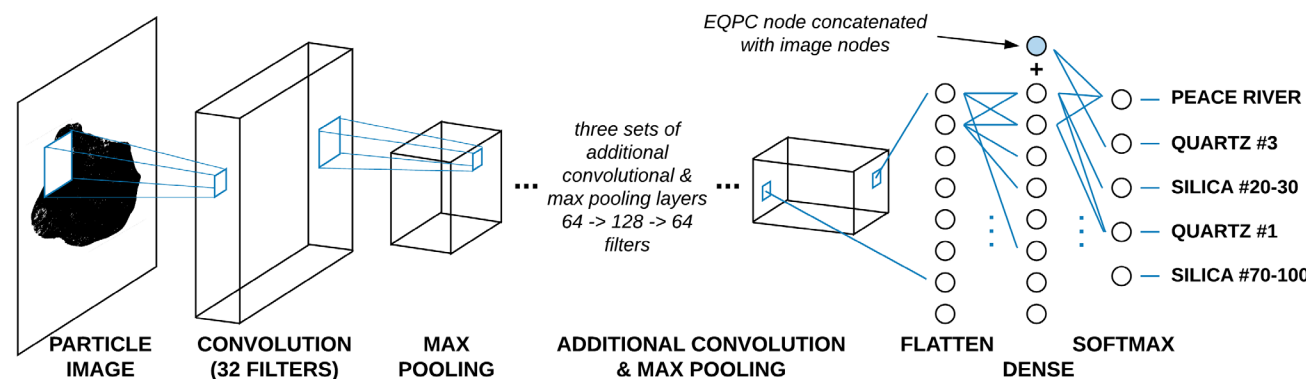


Figure 4. A diagram of the CNN model for particle image classification.

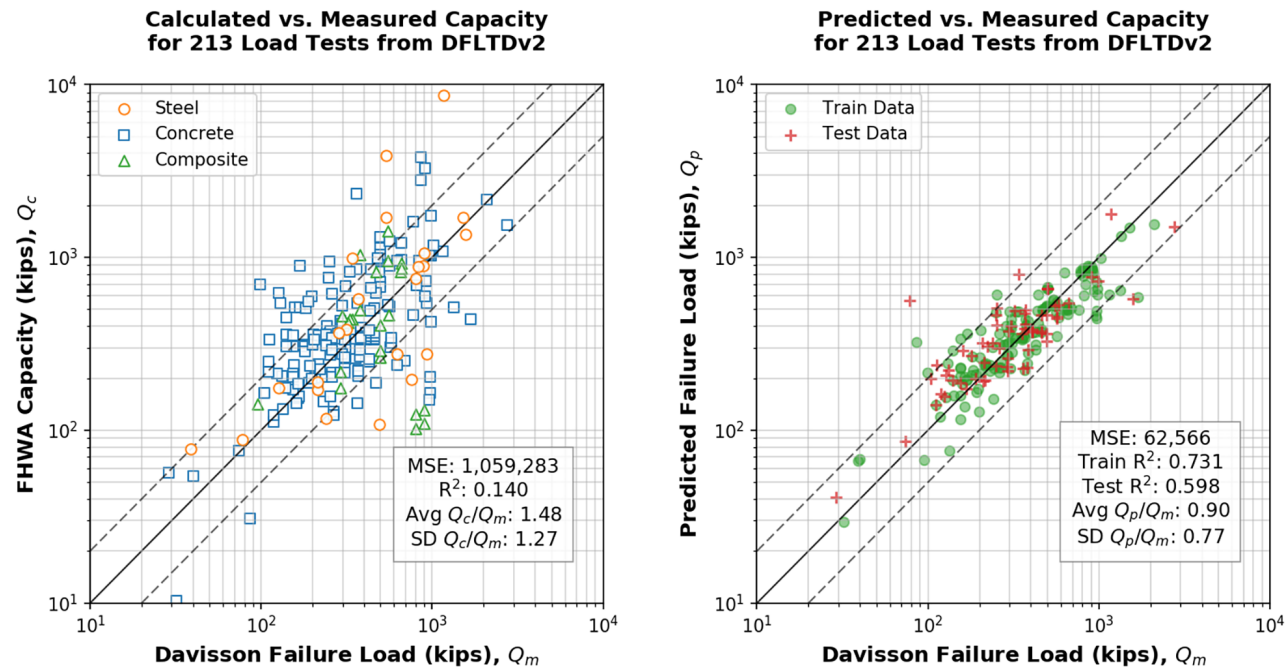


Figure 5. Comparison of calculated and ML-predicted axial pile capacity.

instructions. Some terms frequently linked to ML are *Supervised Learning*, *Unsupervised Learning*, *Semi-supervised Learning*, and *Reinforcement Learning*. These define different groups of ML algorithms. Supervised and unsupervised learning represent the majority of the ML use cases. Supervised learning can be used for classification or regression problems, for example to predict soil types from CPT data (classification) or pile capacity from CPT data (regression). The goal of unsupervised learning is to cluster input data together in meaningful groups by identifying common characteristics among the data.

Although online courses are available, the secret to success with ML is having a good grasp of the fundamental ML concepts and a lot of practice. Be creative, and invent new use cases. For instance, working from the example presented in Figure 3, one can design a classification algorithm that can learn from sieve analysis results and classify soils, certainly faster, and perhaps more reliably, than humans.

Examples

Deep Learning for Soil Classification
In a classification-type ML proof-of-concept study presented at Geo-Congress 2020, we made the case that images of individual soil particles can be used to successfully design and train a Convolutional Neural Network (CNN) model, which is most often used to analyze visual imagery. Figure 4 is a schematic depicting the main components of the CNN, composed of layers of connected neurons with inputs and outputs and learnable weights and biases. Five different types of siliceous sand with different particle shapes were selected for investigation. A training dataset of grading and shape properties of these sands was compiled from over 50,000 images and expanded to 600,000 images with 12 rotations in order to increase prediction accuracy.

The study proved that the building blocks to achieve practical soil classification during site investigation activities are available right now. The training set can be expanded from five to dozens of sand types to cover a wide

range of subsurface conditions. In the future, on-site engineers will be able to use small mobile devices to quickly classify soils. They could possibly offer recommendations simply by capturing a picture of the soil — either by extracting the sample or using a vision cone. More importantly, such a system can aid decision making by eliminating subjectivity and human error.

Predicting Pile Capacity

In a regression-type ML proof-of-concept study presented at IFCEE 2018, we explored the use of ML to predict the axial capacity of piles. A support vector machine (SVM) regression model was designed and trained using 213 load tests curated from FHWA's Deep Foundation Load Test Database v.2 to evaluate the performance of the developed approach against design methods evaluated in FHWA *GEC 12, Design and Construction of Driven Pile Foundations* for the axial geotechnical capacity of single piles in soils. The results of the predictive analysis show

an improvement over the capacities obtained by design methods presented in FHWA *GEC 12*. Perhaps more remarkably, the predictive model outperformed the FHWA pile design method by relying only on seven readily available and easily obtainable features (i.e., soil type, average N , pile material, pile end conditions, cross-sectional area, circumference, and length) as opposed to a laborious and error-prone (when performed manually) design methodology. As shown in Figure 5, the study demonstrates the potential of ML in deep foundation design. For reference, 1:0.5, 1:1 and 1:2 ($Q_c:Q_m$) lines are shown to help illustrate data scatter. Figure 5 shows that the ML-based analysis of axial pile capacity reduced the mean squared error (MSE) by a factor of 17 to 62,566 kips.


What Lies Ahead?

There are several examples from other, more AI-mature fields that geotechnical professionals can learn from. The potential for disruption is real, and those who “get in the game” first will undoubtedly have significant advantage relative to their competitors. When looking at the components of the *AI Trinity* for geotechnical engineering, the algorithms are available, the infrastructure is available, but we lack the data. ImageNet, a massive image database that was made freely available to all, is often credited for many significant advances in image analysis and AI. A similar dataset of geotechnical and foundation data does not exist. This despite the fact that there's an enormous amount of valuable data collected within public and private organizations.

There's no need to reinvent the wheel. To incorporate AI within their normal operations, companies can reach out and work with experts who have experience in building custom AI solutions. AI adoption is a top-down decision. Management must realize the benefits and either rely on cross-disciplined experts or invest in building new data teams.

The first step for geotechnical professionals interested in using data analytics in their practice is to learn the basics of a programming language, preferably Python.

It's important to note that, in the same way that robotic surgery has not replaced surgeons, but rendered surgical procedures far safer, AI will not replace engineers. Rather, it will complement engineering experience with reliable AI predictions and reduced risk. In time, AI will allow for optimized designs with less uncertainty. The corresponding increase in the confidence of our designs will lead to significant savings in design, construction, and maintenance.

Abraham Lincoln is often credited with the notion that the best way to create the future is to create it yourself. Our perspective is heavily influenced by our professional experience and ongoing work at NYU. We encourage readers to explore how *you* can create the future of AI in geotechnical engineering, according to your own vision. 

► **NICK MACHAIRAS, PhD, A.M.ASCE**, is a geotechnical engineering and applied analytics consultant with experience in building custom business and engineering AI solutions, thus minimizing risk and construction costs. He is also a lecturer at Columbia University and New York University, where he teaches graduate courses on modern database systems and machine learning. He can be reached at nickmachairas.com.

► **MAGUED ISKANDER, PhD, PE, F.ASCE**, is professor and chair of the Civil & Urban Engineering Department at New York University's Tandon School of Engineering. A geotechnical engineering educator with 25+ years of experience, his research focuses on physical modelling with transparent soils, polymeric piling, and applications of data analytics in geotechnical engineering. He can be reached at iskander@nyu.edu.