The Accelerating Importance of Data Science in Remediation

by John Horst, Shawn Burnell, Robert J. Stuertzle, Leanne Murdie Austrins and Denny Schanze

Among the most fundamental goals of a remediation practitioner or portfolio manager is to drive progress toward the desired outcomes (protecting human health and the environment first and foremost) in as efficient and timely manner as possible. The ability to do this requires a firm handle on a number of facets:

- Measurable operational goals established for each defined process
- Proactive identification of operational defects or non-compliance
- Anticipation of the need for remedy adjustments to drive performance
- Vision for and understanding of how each project/site contributes to the business strategy and goals

In the past, this has relied on small data sets, collected using manual methods, with collection of the data and analysis separated by months, if not longer. Today, our level of digital maturity within the industry allows for acquisition of much larger data sets, sometimes in near real-time (see our last column on The Rapid Advancement of Environmental Sensors in Remediation; Horst et al. 2022). This, in turn, has changed the nature of analysis. The common view on analytics involves four categories related to their outcome (Cote 2021):

- Descriptive—What happened?
- Diagnostic—Why did it happen?
- Predictive—What might happen?
- Prescriptive—What should we do next?

Tasks farther along the list tend to require more complex analysis but may be high-value efforts because they allow stakeholders to make informed decisions that improve the chances of project success. Historically, descriptive and diagnostic analytics have been the most common and most accessible for the remediation industry. As technology has improved, the ability to shift into predictive and prescriptive analytics has also greatly improved, but it still does not happen at most sites. This is about to change.

In this column, we will focus on how data science can support continuous improvement in remediation portfolio management. To set the stage, we will explore the basics behind data science, followed by the prerequisites that make it possible. We will then examine several example applications of data science in remediation. The first example will examine the optimization of sampling programs through the evaluation of regional variability of analyte requirements, duplicates/trip blanks, productivity, scoping standards and variability. In a second example, an evaluation of the effectiveness of various remedy types across large portfolios is presented.

Data Science—The Basics

At the time the term “Data Science” was coined in the 1960s, there was no way of predicting the truly massive amounts of data that would be generated over the next 50 years, “Data Science” describes the gathering, handling, and interpretation of large data sets (Foote 2021). Since the 1960s, the number of careers in this field has exploded (DiscoverDataScience website 2022). Careers include not only data scientists and analysts, but database developers and administrators (developers handle the design, programming, construction, and implementation of database solutions; administrators safeguard and maintain them); data architects (creates the blueprints for data management systems to integrate, centralize, protect, and maintain data sources); data engineers (focus on the hardware); data mining specialists (use statistical software to help evaluate and model relationships in a data set); and business intelligence analysts (review competitor data and industry trends to establish where the company stands and identify opportunities to improve).

The term “data scientist” can sometimes be used interchangeably with the term “data analyst” but they are not the same. The following is a comparison of the common tasks that each might expect to be responsible for (Coursera 2022):

Data Analyst:

- Collaborating with stakeholders to determine informational needs
- Acquiring data from primary and secondary sources
- Cleaning and reorganizing data for analysis
• Analyzing data to spot trends and patterns that can be translated into actionable insights
• Presenting findings in an easy-to-understand way to inform data-driven decisions

Data Scientist:
• Gathering, cleaning, and processing raw data
• Designing predictive models and machine learning algorithms to mine big data sets
• Developing tools and processes to monitor and analyze data accuracy
• Building data visualization tools, dashboards, and reports
• Writing programs to automate data collection and processing

In short, a data analyst uses existing tools to clean, reorganize and gain insights from the data then effectively communicate those insights to deliver value in business and applied science applications; data scientists use advanced mathematics and computer science to develop the algorithms and tools to facilitate advanced analytics process described above.

There are many different versions of a famous Venn diagram first created by Drew Conway in 2010 to depict the intersection of skills needed by a data scientist. A variation on the more current versions is shown below (Figure 1, Werner 2016). The premise is that to be well rounded and competent, a data scientist should have a thorough understanding of the specific business domain, a strong background in math and statistics, and a strong background in computer science and coding. Having only some of these can support activities other than data science or create “danger zones” where professionals create outputs that they may not understand.

This combination of skills will be more and more in demand as businesses in many different industries, including remediation, have the increasing need to apply data science to make sense of the massive data sets in their domain.

Data Governance—The Prerequisites

There is a lot that goes into handling data that lie “below the waterline”, meaning it is not readily apparent in the visible outputs that are generated. There are dozens of variations of the data analysis iceberg diagram below, all of which are designed to show the level of effort that sits behind / beneath the outputs or products. Much of the work below the waterline in the iceberg figure (Figure 2, which was adapted from holistics.io [Holistics.io website]) is handled by data managers and data engineers. The former (managers) handle the processes of collecting, storing, and applying data in a way that is secure and cost-efficient. Thoughtful data management includes discipline about what is being collected (focusing on data that will drive decisions) and enables organizations to optimize the insights they can obtain from their data, and thereby make data-driven decisions.

Much of the work below the waterline in the iceberg figure (which was adapted from holistics.io website: Woon 2019). In today’s digital economy, data management is more important than ever—a data set is a kind of capital, and strong management practices and solid management systems are an important factor in the growth of these assets. By comparison, the engineers deal with the structural applications of data management. A data engineer creates the data warehouses and data lakes and builds the pipelines that convey and transform data for further analysis or visualization. Proper data management and engineering ensure reliable and consistent data for further analytics and are prerequisites for efficient data science.

When considering the quality of data, we have all heard the mantra “garbage in, garbage out.” Before we attempt to derive any insights from a data set (or sets), the quality of the underlying information must be assured. This is the focus of data governance. Here, the focus is on standardizing the way that data are collected and stored. The attention to this foundational prerequisite has increased in recent years as the size of our data sets has continued to balloon.

In remediation, commercial analytical data received from a laboratory has traditionally been the most well-governed data. However, the efforts have been far from perfect. Data
set owners typically defer the methods for data handling to each individual supplier. Not only does this make it difficult for the owners to access all of the data for a portfolio (or even for an individual site) but it introduces multiple variables defined by the practices of each supplier. This situation is not necessarily improved by adopting structured databases that are hosted by the owners, as the handling differences can still carry over. This issue spells disaster for high quality analytics or data science aimed at supporting automated outputs.

The rise in importance of other types of data in the overall spectrum of site information has only increased in complexity. In addition to data collected for a project or program, publicly available data are also being incorporated into the knowledge base associated with various remediation programs. This can include data from governmental agencies such as the USGS, EPA, NOAA, and others. As the types of data that are of interest to a remediation project continue to diversify, it has become clear that there is value in integrating the multiple different disparate data sources and also emphasizing metadata to be able to correlate across databases. This can lead to a more holistic understanding of a problem, especially where intelligence tools are used to help show the associations and patterns across datasets in ways that would not have been possible before. As more datasets are included in projects, the quality of the underlying data becomes more important, and the importance of consistent data governance increases.

In our last column we mentioned the evolution of the relational database model that records data in a rigid structure of rows and tables with logical links between them, to a non-relational format that offers greater flexibility with modern data sets, allowing different structures to sit alongside each other, permitting easier lateral scaling (Horst et al. 2022). As options like this, and their flexible pairing to other tools (sensors and other data collection devices as well as data visualization tools), become more mainstream, the comparatively rigid structured database options that have been sold commercially for remediation will become less relevant, despite the attempt to link them more closely to field data collection and visualization in a way that would capture the data life cycle.

Selecting Data Science Projects

With more sources of data available and mined, the opportunities for leveraging data science and analytics are increasing rapidly. The point has been reached that there now needs to be some rigor applied when selecting opportunities for data science projects—to avoid low value/high effort situations.

One important question to start with is: What is the value? Obviously, efficiency is the easiest to justify and understand—i.e., supporting a reduced timeframe and cost to achieve the remedial goal. However, there can be other sources of value that are perhaps viewed as softer. Things like safety, sustainability, and productivity. These benefits are harder to quantify, but they are important and they are all linked. Issues with safety can result in unacceptable risks to stakeholders and workers, so anything that can reduce the risk profile of a project is beneficial. Productivity can improve safety. It can also improve sustainability. Most organizations are now committed to helping transition to a net zero carbon world. Remediation projects can contribute to that goal—and data science can help with all of these aims.

As an example of this, we have two case studies that focus on groundwater sampling. Routine groundwater sampling is often a significant portion (10–50%) of the lifecycle cost of an average remediation project. Using data science, we can examine data in a way that helps:

- Focus sampling where it ensures compliance and manages risk,
- Eliminate the collection of data that are not used in decision making,

These decision aids can optimize the sampling program to improve its direct relevance to compliance, risk management, and decision making. They can also reduce the overall carbon footprint of the program and aid with the focus of all stakeholders on the data that matters, helping to drive progress.

Example Application #1—Groundwater Sampling QA/QC

In our first example, the team involved examined the portion of groundwater samples across a large Operations and Maintenance portfolio that were collected as part of routine quality assurance/quality control (QA/QC)—namely duplicate samples and blanks (field, trip, and equipment). For remediation projects under its jurisdiction, the USEPA requires duplicate samples be collected at a rate of 10% of the overall number of samples. This requirement was introduced over 30 years ago, when laboratory technology and data management were less reliable, to serve as an alert to questionable analytical accuracy through either poor method management or cross contamination. Beyond simple comparisons, statistical analyses related to these samples are rarely undertaken. Over the last 30 years, laboratory procedures and analytical methodologies along with field procedures have been standardized and improved, resulting in a significantly improved quality of analytical data.

For the subject portfolio, the data science effort was aimed at gathering, cleaning, and processing the raw data for 512,552 points of analytics data completed over a 10 year period working with 16 various laboratories. A model was created to run the comparative analysis, and utilize data visualization tools to convey the results. Additional data moving forward were analyzed in a consistent fashion. The team found that of 512,552 analyses there was only a 0.9% rate of detection (88 samples) of any analyte in the QA/QC blank samples. Of those 88 detections, 65 were associated with acetone or methylene chloride, which are common laboratory background contaminants and were also two orders of magnitude below the regulatory standard in the highest detection. Of the remaining 23 detections, only one was determined to have added value to the project. A contaminant of interest at the site was detected in a field blank at 6.5 μg/L compared to a regulatory standard of 5 μg/L. None of the rest of the blank results were used to change any decision making or prompt resampling because they were negated by the “5x rule” (the detected concentrations in the samples were less than 5 times the blank concentration and
were therefore reported as a non-detectable (ND)). So of the 512,552 analyses, only 1 or 0.01% resulted in a changed result to a sample (Figure 3).

Based on the data science applied to this very large set of data, the team determined that laboratory improvements now make it possible to achieve desired levels of quality assurance and quality control (QAQC) with fewer duplicate samples or blanks, as they relate to routine groundwater monitoring. In a similar manner, the statistical methods for evaluating groundwater monitoring data have also progressed. In 2009, the USEPA issued new groundwater statistical guidance (USEPA 2009). In this guidance, new strategies were presented for controlling the statistical false positive rate (SWFPR) in which resampling is involved when results are outside of a calculated statistical range for a given constituent. The SWFPR is the rate at which a statistical test will incorrectly identify an exceedance of these predetermined interval estimators. Because these methods (and this guidance) allow for resampling, statistical methodology affords some of the same protection that was once only achievable by the inclusion of blanks and duplicate samples. This exercise demonstrates that by using analytics, the traditional thinking related to blank samples could adapt a more “fit for risk” approach. By using analytics, it would be possible to improve our QA/QC, reduce the carbon footprint and effort associated with collecting these samples and ultimately achieve the desired results with no decrease in reliability.

Example Application #2—Groundwater Sampling Program Optimization

In our second example, the team involved desired to understand whether or not sites across a large portfolio were being over sampled relative to similar sites in similar jurisdictions. A machine learning model was trained to predict which site characteristics resulted in a site being sampled more or less than other similar sites. This training was based on:

- Site information: city, state, and regulatory agency
- Analyte concentrations
- Well information: groundwater elevation, well depth and sampling method

Physical site characteristics were not included in the model, because the data were not readily available. Also, for the purposes of this screening tool and to demonstrate the machine learning steps and principles, a more complex analysis was neither necessary nor practical. The dataset used to train a “model” included 3 years of sampling across hundreds of sites in the United States. The model consisted of an evaluation of sampling frequency, an average number of monitoring wells sampled per event (or year), and analytes per well for each facility type. The team was able to distill the top three parameters in the algorithm most closely correlated to the number of samples collected:

- mean concentration
- max concentration
- state

This seemed to reinforce what many practitioners that work across many states and regulatory jurisdictions suspected; that there can be a wide range in requirements depending on regulatory jurisdiction. Being able to approach regulators with data showing ranges of sample numbers on similar sites provides opportunities to move beyond exhaustive site-by-site sampling plans that require years to optimize. This approach can lead to better site management and less waste. The term “optimization” can refer to the correction of undersampling as well as oversampling, based on site risks anticipated by the model.

The data science to accomplish this made use of machine learning algorithms and SHapley Additive exPlanations (SHAP). SHAP values are used to increase transparency and interpretability of machine learning models. A SHAP value of one would mean that the feature being tested returned the baseline prediction of the model. A SHAP value greater than one would mean that the feature being interrogated by the model (concentration, state, etc.) would cause the model to under predict the amount of oversampling or undersampling.

Figure 4 shows the SHAP value results associated with the maximum concentration. In this case the very low concentrations have a wide range of SHAP values (−4 to 3) so these concentrations (<500 μg/L) do not really help the model to predict if a site will be normally sampled or over sampled. The moderate concentrations (1000–5000 μg/L)
drives the model to predict that a site will likely be oversampled. The highest concentrations (>5000 μg/L) had less of an effect on the model and while it still pushes the model to predict that a site will be oversampled, it is not with as much emphasis as the moderate concentrations (higher SHAP values).

Figure 5 shows the SHAP value results as they relate to state. There was only a positive SHAP value for five states. That suggests that in most states the model was less affected by state than other features included in the model. This indicates that while we think of states being important, it may be subordinate to other factors that lead to oversampling or undersampling. So while “state” was the third most important factor, maximum and mean concentrations were significantly more important in determining if a site was going to be over or under sampled. Using the model to look at a portfolio of projects can help pinpoint which sites might have opportunities to reduce sampling and also sites that might be under sampled compared to similar sites, which might present a project level risk.

Example Application #3—Remedy Optimization

TISR® (Thermal In-Situ Sustainable Remediation) is a sustainable remediation technology using solar water heaters to increase the subsurface temperature to accelerate natural degradation and/or phase partitioning (Horst et al. 2018). As part of a pilot implementation at an oil terminal in the Netherlands, there was a desire to understand how seasonal variables and site conditions would affect the operation of the TISR® system.

To enable real time insights in the remedial and system performance, data science methods were introduced. Multiple data sources were combined to create a monitoring dashboard. These data sources included:

- Digital field forms
- Automated data exchange between field, laboratory and office utilizing
- Process controls and thermocouple data from the on-site SCADA system
- Weather and solar radiation data

The different data sources were integrated using APIs (application programming interfaces), which gather the data from a range of sensors into an MS Azure cloud system, and then processed and presented the information using interactive Power BI Dashboards (Figure 6). This setup allowed for efficient data-driven decision making, as the data could be evaluated in real time and changes to process parameters could be immediately implemented. Furthermore, the dashboards improved client communication and engagement since all data were at hand during meetings, whether in person or online.

To increase efficiency even more, the team is now applying additional data science techniques in the form of machine learning, so that the data monitoring will allow the system to suggest changes for efficient operation, getting closer to prescriptive analytics. The multi-year ambition is an autonomous AI driven operating process, keeping manual operation and maintenance at a minimum. This is an amazing feat to consider. Most automated system interfaces in the remediation industry are merely translations of alarms and measurements that still require significant oversight and intervention from an operator. The use of data science may allow some types of remediation to achieve a future operational design that requires minimal human intervention and achieve more fully autonomous operation. It can also respond in real time to the need for operational changes which can drive faster progress toward the end goal for treatment. As we get closer to autonomous operation, the constraints on which changes can be made autonomously and which cannot, will be a new topic for stakeholders to discuss.

What Is Next?

As an industry, we are just starting to explore what we can accomplish using data science. As shown by the examples highlighted here, machine learning and artificial intelligence are finding their way into our data review and system operations. Gaining momentum requires more data sources to have established governance that standardizes their
collection and management. This will introduce larger data sets and more opportunities for data science and advanced analytics that drive not only technical efficiency, but also improve health and safety, and reduce the carbon footprint. As technical practitioners continue to innovate leveraging data in new ways and business critical benchmarks continue to be identified and optimized; the presence of analytics in day-to-day business will increase. Like the TISR example above, an increased use of sensors will support this transition. The sensors will provide more real time data, on which machine learning and AI enabled operational support will evolve beyond what have been used to in our industry. This sort of operational support can result in lower life cycle costs for operation of engineered treatment systems through the combination of more timely maintenance (avoiding the cascading complications and cost of component failure) along with more accurate forecasting. To this end, advancements using AI that we are already seeing in the asset management and municipal O&M fields (Steel 2018) have the potential to be adapted to help manage the mechanical components of remediation systems particularly for remedies that have no foreseeable end date.

As we consider the above, the use of data science through the application of AI and machine learning will support the generation of predictive maintenance schedules for equipment, ensure compliance, and optimize the restoration dollar spent. We can envision the focus then shifting from improving the value of each dollar spent to minimizing expenditures. Once this operating model is more ubiquitous across the industry, it will conceivably drive more standardized pricing and a procurement model for restoration services shifted toward unit rates informed by benchmarks with incentives tied to productivity. This is a win-win for the client and consultant. Clients will have better cost certainty and the consultant is incentivized to increase productivity.

This sort of arrangement can also be applied beyond remedial system operation to health and safety and project/portfolio management. Predictive analytics for health and safety can have industry wide benefits. With behavior-based safety (BBS) systems the focus has been on leading indicators and the personal engagement of individuals in the project to encourage safe behavior that avoids incidents. By using digital tools to increase proactive engagement with data related to safety, industry has seen a benefit in the ability to reduce injuries (Roberts 2021). By including other data sources such as weather, project financials & schedule, facility type and season; the opportunities to apply machine learning grow to the extent that project teams can be proactively alerted of the most probable risks based on their pending project work, or even alert staff in the field of an impending risk.

As our industry matures in the use of data science we can envision the scenario where data are shared between stakeholders and their consultants, subcontractors and vendors, reducing duplication. Geospatial, site-specific, technical and financial data are often needed by multiple parties involved with a given project. The more efficiently the data can be shared between these parties the better the coordination (which can help drive progress), and the lower the net overall cost and energy consumption associated with storing and managing the data. This in turn will make accessibility and management paradigms related to project data a more...
central part of how project decisions are made and how we all work together.

As part of a trend everyone can relate to, recent graduates are more digitally savvy than the generations that preceded them, and our workforce needs are demanding these skills (more people are in intersecting parts of the Venn diagram shown above). This shift in the skills we value will only accelerate along with the use of data science. Enabling this new workforce to continue discovering the digital innovations that can advance the state of our practice requires deliberate approaches at an organizational level to manage “citizen development.” The term citizen development (Gartner website 2022) applies to cases in which digital savvy team members that are not part of IT, but are using tools that are enabled by IT, are creating or testing capabilities designed to support a specific project that may ultimately have the potential to be scaled across an enterprise of hundreds or thousands of users. Innovation culture and systems will continue to evolve to include citizen development of digital tools.

We are headed into an exciting time—where “George Jetson” level technology (the kind of technology that we might have only imagined in a fictional futuristic cartoon) is becoming reality. And that new reality will continue to shift our industry to one of greater and greater value creation.

References

Cote, C. 4 Types of Data Analytics to Improve Decision Making. 19 October 2021. HBS Online, Business Insights Blog. https://online.hbs.edu/blog/post/4-types-of-data-analysis. (accessed June 3, 2022).

Coursera. Data Analyst vs. Data Scientist: What’s the Difference? 26 May, 2022. Website Articles. https://www.coursera.org/articles/data-analyst-vs-data-scientist-whats-the-difference. (accessed June 3, 2022).

DiscoverDataScience.org website. Data Science Career Guide Overview. https://www.discoverdatascience.org/career-information/. (accessed June 3, 2022).

Foote, K.D. A Brief History of Data Science. 16 October 2021. Dataversity Data Education Topics. https://www.dataversity.net/brief-history-data-science/. (accessed June 3, 2022).

Gartner.com website. Gartner Glossary. https://www.gartner.com/en/information-technology/glossary/citizen-developer. (accessed June 17, 2022).

Horst, J., C. Flanders, M. Klemmer, D.S. Randhawa, and D. Rosso. 2018. Low-temperature thermal remediation: gaining traction as a green remedial alternative. *Ground Water Monitoring and Remediation* 38, no. 3: 18–27.

Horst, J., N. Welty, J. Gallegos, D. Dunham, and A. Adams. 2022. The rapid advancement of environmental sensors in remediation. *Ground Water Monitoring and Remediation* 42, no. 2: 12–19.

Roberts, H. Why You Should Include Behavior-Based Safety in your Safety Management Program. July 13, 2021. EHSToday Online. https://www.ehstoday.com/safety-leadership/article/21169471/why-you-should-include-behaviorbased-safety-in-your-safety-management-program. (accessed June 9, 2022).

Steel, W. Machine Learning in Water Asset Management. July 1, 2018. Waterworld.com website. https://www.waterworld.com/water-utility-management/asset-management/article/16190079/machine-learning-in-water-asset-management. (accessed June 21, 2022).

USEPA. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities – Unified Guidance. March 2009. EPA 530-F-09-020. https://archive.epa.gov/epawaste/hazard/web/pdf/unified-guid-fs.pdf.

Werner, G. Data Science Venn Diagram V2.0. 15 September 2016. Illumidesk blog post. https://medium.illumidesk.com/data-science-venn-diagram-v-2-0-c0250319e819. (accessed June 3, 2022).

Woon, V. 2019 Make Tableau Work Better for You with Holistics. Holistics Blog post March 11, 2019. https://www.holistics.io/blog/make-tableau-work-better-for-you-with-holistics/. (accessed June 21, 2022).

Biographical Sketches

**John Horst**, P.E., corresponding author, is Global Strategy & Development Officer at Arcadis, 10 Friends Lane, Suite 100, Newtown, PA 18940; (267) 685-1800; john.horst@arcadis.com

**Shawn Burnell**, is Market Sector Leader with Arcadis, 482 Congress Street, Suite 501, Portland, ME 04101; (207) 613-8416; shawn.burnell@arcadis.com

**Robert J. Stuetzle**, P.Geo, is Remediation Manager at Dow, Inc., 633 Washington Ave., 1803 Building Midland, MI 48674; rjstuetzle@dow.com

**Leanne Austrins**, is Senior Remediation Manager at Dow, Inc., 633 Washington Ave., 1803 Building Midland, MI 48674; laustrins1@dow.com

**Denny Schanze**, is a Senior Consultant with Arcadis, Beaulieustraat 22, 6814 DV Arnhem, The Netherlands; denny.schanze@arcadis.com