Space is big. Really, really big. And the data involved with space exploration and the related sciences that are in NASA’s purview are … astronomical (sorry NASA—hack joke equivalent to the taxidermy/taxonomy references that I hear). Since its founding in 1958, NASA has led the US government’s exploration of space. Its contributions have extended beyond the government missions it supports to advances that have benefited other parts of the federal government and many industries in the private sector. Because of its unique requirements for data collection and analysis, NASA has been on the forefront of many innovations in data and computational science, and related IT.

The Data in Dust Devil Details
One area of interest for the agency is real-time event detection. The problem is, what is real time on Mars? When NASA sends spacecraft to Mars and other distant locations, they are programmed to complete a variety of specified tasks. However, to make the best use of their resources, these vehicles ideally would have a certain amount of autonomy, because planning for every contingency is difficult or impossible. In studies of the surface of Mars, one phenomenon that occurs at frequent but unpredictable intervals is the appearance of dust devils. Because of the time delay in speed-of-light communications reaching from Mars to Earth, NASA scientists are unable to know in real time when such an event takes place. For many years, common wisdom said that the opportunity to redirect resources on the surface to study dynamic phenomena such as dust devils as they occurred was out of reach.

By putting change-detection capability on the Mars Rover Opportunity, however, real-time detection became possible. This capability allows the Rover to suspend whatever observation plan was in place and focus on the dust devil. When scientists back on Earth become aware of the event, they already have a complete information package of it in the form of the video sequence.

“The only way to have this kind of autonomous response is to put enough computational capability and analytic capability on the Rover itself,” says Richard Doyle, program manager for information data science at NASA, “and authorize—with scientists’ involvement and blessing—that the system would have this kind of autonomous response.” There are many other variations of this type of activity at NASA, but this is a good example of how the organization is using analytics in an innovative way, and it is paying off in the scientific community. “Up to now, our scientists have assumed they weren’t going to be able to capture and track remote transient events,” says Doyle, “but with the right computational capability and system architecture, we can do it.”

Programming Triggers for Scenarios of Interest
To accomplish this goal, the Rover has to interrupt what it’s doing, which it does via onboard reasoning about capabilities, resources, and priorities. But what are the triggers? It would seem that this type of response would have to be exhaustively programmed. The challenge is to generalize a trigger so that it can apply in a multitude
of scenarios. This paradigm—event detection and response—requires intense interdisciplinary collaboration. “Atmospheric scientists, planetary geologists, and other experts sit down with us and explain what events would be of interest,” says Doyle. “It should be something that is important enough to interrupt ongoing surface mapping and imaging, and attend to a previously identified science opportunity.”

At this point, the computer programmers who develop algorithms devise a technique that is robust enough to detect an event of significant interest. The architecture needs to be developed with an onboard planner to allocate capabilities and resources to the event as it’s occurring. Many actions might be required for the response. For example, the camera needs to be pointed in the right direction, power needs to be available, and the constraints of the Rover to operate safely need to be considered. Such important details are part of the event detection and response framework.

**Artificial Intelligence and Machine Learning**

Machine learning is an essential component of this process. In the training phase, experts identify examples of events of interest. Through a variety of techniques, a generalized pattern is determined to enable the computer to then find additional examples of the same phenomenon or event in subsequent data. This process is referred to as supervised learning because it is guided by an expert who provides an initial set of representative examples.

The number of training examples required to achieve success depends on the data, how it varies across examples, and how noisy it is. “Ideally, the system would have hundreds of examples,” says Doyle. “But solving this kind of problem can be very context-dependent, depending on what’s going on with the dataset.” It can be a tedious exercise on the part of scientists or subject matter experts. “But when they see that it works, they can be quite impressed, and they want more, especially when they see how the software can function as a proxy for their interests,” adds Doyle.

**Unsupervised Learning and Outlier Detection**

Another approach to machine learning is unsupervised learning. An example is the process applied to classify information in a set of unstructured documents. Analyses can be done to find the natural clustering and divisions in the data. Unsupervised learning can look for inherent patterns in data and group those patterns. The approach can also look for data that falls outside of a particular pattern—the identification of anomalies or outliers. In this case, identification of outliers can serve as a form of event detection. You might not know what those outliers look like ahead of time, but the software can surface these for a human to review. “The question of what part of the data is most unlike the rest can actually be framed in a very robust computational manner,” says Doyle. “You are asking a more general question. You can get some very interesting results without having to perform explicit training that specifically identifies the outliers.”

**Real-Time Detection and Earth-Observing Spacecraft**

Other examples of real-time event detection can be found closer to home than Mars. “A significant portion of NASA’s work involves Earth-observing spacecraft,” says Doyle. “In those cases, you don’t have light-time delays, but you still want a quick response.” Examples include natural disasters such as forest fires, sea ice breakup, floods, and volcanic plumes. “You want to be able to detect them from orbit very reliably, and then initiate an appropriate alert in as timely a way as possible,” Doyle says. “The alert needs to be automated because otherwise it might be hours or days before a scientist looks at the data, and in the event of a natural disaster, you need to do better than that.”

In a recent study by the US National Academy of Sciences, Frontiers in Massive Data Analysis, one of the conclusions was that integrating analytics across the entire sequence of the data life cycle is critical. These stages include the creation, capture, processing, management, and distribution of data. “When an image is created, the features should be analyzed in order to better inform the search engines. Then scientists can begin to search on those features,” says Daniel Crichton, coauthor of the report. Crichton is the NASA Jet Propulsion Laboratory’s (JPL’s) program manager for data science and leads the Center for Data Science and Technology, a joint initiative with Caltech. “Techniques such as computational algorithms, statistical algorithms, and machine learning need to be applied to the whole life cycle, from considerations on board all the way to integration of distributed data from scientific archives.”

**Big Data, Integration, and Predictive Analytics**

NASA builds large data stores as a basic product of its missions, and recognizes the value of being able to integrate the data. “That is a very difficult problem,” says Crichton, “but in many cases, if we want to understand a problem, we need to consider many data sources and integrate them.” For example, to understand the drought in central California, NASA might need to look at fusing together many types of data from multiple measurements across satellite, airborne, and...
Data Analytics

Ground-based sensors to better understand the water dynamics. Computational methods are needed to integrate and reduce the data to understand what it means.

Inferences are made that provide results at a certain confidence level, which is a step toward predictive analytics. “For example, we want to be able to understand the dynamics of climate change, such as water availability and the total water balance for a specific region.” Tools now exist to help scientists in building better models and leveraging many of the measurements from NASA missions. In the past, they could do much of the analysis within their own local environment, but because of the massive volumes of data, this now requires new computational approaches.

The total capture of Earth science mission data is now about 12 petabytes, but in the next five years, the volume is expected to grow by an order of magnitude. “We need to be developing intelligent algorithms to extract data from larger datasets,” says Crichton, “so we will need to be able to apply more automated methods of data reduction.” A major challenge now is to work with a big data infrastructure to download subsets of data, but this brings on another problem: the higher level of uncertainty that results from taking a small sample.

“It becomes harder to reproduce some results because a lot of the knowledge and understanding may be in the mind of the scientist that did the analysis,” says Crichton. NASA has enormous amounts of archival material and enormous amounts of expertise in the minds of scientists and engineers. “We are trying to systematize and automate the analysis so results can be reproduced on demand, rather than having to rely on humans to go back and reproduce it.” For example, if someone is reading a paper and has questions about the results, is it possible to go back and validate the results, or do the results change if more data is incorporated into the analysis? In this case, the model might need to be adjusted.

Data Architecture and Governance
Reproducibility depends on being able to find the sources of information, which in turn depends on having a system for organizing it. NASA has a long history of building sophisticated data models and taxonomies that can support its search and analysis needs. “The Planetary Data System used for many years is what I would describe as an implicit data architecture,” says Crichton. “Data dictionaries were developed, and there were forms of governance to manage change. Metadata was defined that was needed to contextualize and describe the observational data that we were capturing and archiving, but it was difficult to systematically link data architectures to software systems as the sophistication of data and software evolved. In the last five years, we have turned it into an explicit data architecture, recognizing the importance of ensuring we can continue to capture the diversity of data from NASA solar system missions by linking the data architecture and software systems so they can evolve in a model-driven architecture.”

NASA has now developed an ontology as an information model for describing planetary science data that includes the mission, the instruments, the observations, and the types and structure of the data. The ontology governs the definition of the data and is architected to integrate with a scalable big data system so the software can adapt as the ontology changes. Through these efforts, NASA has ensured that for future missions, the data can be generated and validated using the ontology, and ultimately shared and used across international missions. “We sat down with the scientists to understand what knowledge we should be capturing from the data for their particular discipline,” says Crichton. “We encode it into the ontology and use that to derive metadata structures.”

A governing board makes any required changes to the ontology. To manage the ontology, NASA uses an open source software product called Protégé.

AI, Machine Learning, and Analytics

JPL has established a joint initiative with Caltech to perform research in data science and data-driven discovery. JPL, via the Center for Data Science and Technology, brings depth in data architecture, the data life-cycle model, analytics, and experience deploying large-scale distributed data systems. Caltech, via the Center for Data-Driven Discovery, led by George Djorgovski, brings depth in basic research, visualization, discovery and analytic techniques, and several science disciplines, such as astrophysics and biology. This interdisciplinary team includes experts in ontologies, visualization, computer algorithms, cyberinfrastructures, artificial intelligence (AI), and machine learning. “All this comes together,” says Crichton. “The models, integration, computational algorithms. We then can begin to understand what can be encoded in future systems to provide more analytics across the entire data life cycle.”

The Caltech-JPL team held a joint Summer School on Big Data Analytics in September 2014. The course was offered as a massive open online course (MOOC) and was conducted over nine days, with instructors available to interact with students in chat-room-like settings. The material remains available.
online (see https://www.coursera.org/course/bigdataschool), and at last count, more than 16,000 students worldwide have taken the course.

The computational infrastructure developed at NASA is being applied to disciplines outside that organization. For example, in the field of medicine, different formats are used by different hospitals. When one hospital acquires another, the acronyms and protocols need to be disambiguated to make the analyses comparable. A goal is to integrate a clinical support system with test diagnostic mechanisms so that physicians and providers can understand the latest in evidence-based medicine.

NASAs is shifting from an era in which it used to capture and archive data to one in which data is analyzed extensively, at every phase of its life cycle. “We are introducing analysis at all stages, and doing it in a much more agile manner,” says Crichton. “This is the only way we can keep pace and maintain our performance in the sciences.”

Acknowledgments
This article is based on extensive interviews with Richard J. Doyle, program manager in the Information and Data Science Office of NASA’s Jet Propulsion Laboratory at the California Institute of Technology, and Daniel J. Crichton, leader at the Center for Data Science and Technology, and program manager for data systems and technology at NASA’s Jet Propulsion Laboratory.

Seth Earley is CEO of Earley Information Science (www.earley.com). He’s an expert in knowledge processes, enterprise data architecture, and customer experience management strategies. His interests include customer experience analytics, knowledge management, structured and unstructured data systems and strategy, and machine learning. Contact him at seth@earley.com.