By Stephen Swoyer

With all of the talk about the breathtaking promise of predictive analytics, it’s easy to lose sight of an inconvenient truth: predictive analytics is hard.

Industry experts have no doubt: predictive analytics holds tremendous benefits, but it’s really, really hard work.

You can quantify its difficulty using any of several metrics—algorithmic complexity; the worldwide dearth of skilled data scientists; or the insufficiency of existing IT infrastructure, particularly for preparing and transforming data.

The inconvenient truth remains: predictive analytics is hard. “There are so many tests to detect traps you might fall into, and none of them is built into the tools because they are dependent on the data,” says industry veteran Mark Madsen, a principal with consultancy Third Nature Inc. He says that predictive or advanced analytics is so difficult that “a layman such as myself shouldn’t be allowed anywhere near most analytic software. The arena of mathematical analysis of data is a deep hole. I wonder how much can really be separated into easy-to-use tools and applications for specific problems.”

Dave Henry, senior vice president of enterprise solutions with business intelligence (BI) and analytics specialist Pentaho Inc., doesn’t disagree with this assessment.

Predictive analytics, he concedes, is hard. At the same time, Henry stresses, it isn’t (or doesn’t have to be) prohibitively hard. He draws a comparison to the developing data warehousing market in the late 1980s, when the core concepts, techniques, and technology of data warehousing were highly specialized and (compared to today) poorly understood.

This didn’t deter organizations from investing in DW technologies, Henry points out, because the potential for high-value ROI was both immediate and unmistakable. Early DW adopters weren’t trying to build out complete BI practices overnight; nor did they expect to immediately be able to embed data warehouse-driven BI across all of their business processes.

They were looking to address long-standing, and seemingly intractable, business problems.

Henry foresees a similar uptake model for predictive analytics—that many customers will adopt solutions that address specific use cases.

“If you just pick a specific problem and you focus on that, it’s less labor- and expertise-intensive. There’s always going to be a high-end data mining vendor community that’s going to go bleeding edge, and that’s important, but for these more straightforward use cases, [such sophistication] isn’t necessary,” he notes.

Henry cites the use of predictive analytics in combination with some streaming or sensor data as a relatively straightforward use case. Take the “SMART” sensor data generated by hard disk drives; Henry says that storage vendors can use Pentaho Predictive Analytics to predict imminent disk or hardware failure. Although embedded diagnostic software generates vast amounts of data, it’s highly structured and the data quality tends to be very good. This makes it feasible to create models that have good predictive power. They don’t have to be perfect; they just have to provide a reasonable economic ROI, Henry points out.

The SMART case also demonstrates why predictive analytics isn’t a turn-key proposition. Microsoft Corp. includes sophisticated data mining algorithms with its SQL Server Analysis Services (SSAS), for example, but this doesn’t mean that SSAS-powered predictive analytics practices are popping up everywhere.

In addition to the algorithms, intuitive, powerful, visual environments are needed for designing predictive models. Pentaho has these capabilities, Henry says; its Pentaho Predictive Analytics is based on the open source Waikato Environment for Knowledge Analysis (WEKA) library. WEKA gives Pentaho a rich library of visualization tools and algorithms, as well as a predictive modeling workbench. On the other hand, he says, WEKA is just part of the equation; the bulk of any data scientist’s time is spent preparing data so it is suitable for the kinds of algorithms to be used. That’s why Pentaho positions itself as a full-blown platform for data mining and analytics,
Henry says: it includes an enterprise data integration (DI) component based on the Kettle open source ETL project.

Take the SMART sensor data use case. Data of this kind includes hundreds of potentially extraneous attributes; before it can be consumed by either a conventional data mining tool or handed off to Hadoop, it first needs to be cleansed and transformed.

“If you have an ETL process that breaks that apart and ... gets that ready as a structure to be fed into the model, and have the model read that data and append a score to it, then in a downstream process you’re able to route the scored records based on thresholds. We can do that today: we can do that outside or inside Hadoop.”

The use of plain vanilla Hadoop for big data analytics isn’t a turn-key proposition. Doing anything on Hadoop requires highly specific expertise—starting with programming proficiency in Java, Pig, Python, HiveQL, and other languages. Pentaho Predictive Analytics generates its predictive models in a Hadoop-friendly format: Java.

“This gives us an ability to operationalize these [data mining] algorithms, package them, or put them together in solutions that people can deploy operationally. In our case, the reason we can do this is that the models we generate using WEKA—the math [of these models]—is literally stored as a Java object on disk,” says Henry.

“For [use cases] that involve huge volumes of sensor data, this kind of analysis is tailor-made for Hadoop, too. We’re working with a lot of companies that have storage array technologies. They’re putting pretty expensive devices in the field. In many cases, they’re OEM providers of storage equipment and the stuff goes out in the field en masse. What they want to do is to start proactively predicting failure. By having a database of diagnostic information that they know is associated with devices that have failed, they can use that to train models ... that will look at device data and apply those regression models and generate some kind of probability of ‘Do we need an intervention?’”

It’s a virtuous cycle, Henry says. “As the models improve, we’ll see more predictive intelligence in the devices themselves. Lightweight Java-based frameworks such as WEKA are good for this scenario, too.”

Stephen Swoyer is a contributing editor for TDWI.