DATA SCIENTISTS AREN’T DOMAIN EXPERTS

STIJN VIAENE
VLERICK BUSINESS SCHOOL & KU LEUVEN
DATA SCIENTISTS AREN’T DOMAIN EXPERTS

Citation Information:

Data Scientists aren’t Domain Experts

"Data science" has been around for a while, but recently, there's been renewed interest in the term. Popularized by a new breed of gurus groomed by big data companies such as Amazon, Google, and LinkedIn, its reappearance signals the next step in the coming of age of what used to be called "business analytics." Data-based decision making is becoming a competitive necessity,¹ and Thomas H. Davenport and D.J. Patil have helped fuel business interest by saying that data scientists have "the sexiest job of the 21st century."²

Although data science isn't an academic discipline or established profession (yet), scientists, technologists, engineers, statisticians, and mathematicians are joining forces with big data business to carve out their turf. The days of the passionate, well-meaning amateur seem to be numbered. Data science is now being benchmarked against practices employed by highly skilled information age professionals, who can make big money using scientific methods to discover knowledge in data.

However, with all the hype around data science, expectations are being blown out of proportion. For data science to really work, you need a multiskilled team (not just data scientists) and projects (not just data experiments). Furthermore, to generate business value, you need to connect the world of data scientists to that of domain experts.

Two Worlds Brought Together

In a recent interview, Digvijay Lamba, a distinguished architect at Walmart Labs, said the following:³

What's happening is there are domain experts — buyers, merchandisers, product managers and others [who] have worked in retail for years and years — these people know the market really well.... They throw these ideas over the wall to data scientists, who go through the data and come up with these brilliant ideas to answer questions. But there is a wall there. The data scientists are not domain experts.... What we want to do is break down the walls.

Lamba's diagnosis is spot on. Data scientists are separated from domain experts, and the challenge is to bring the two together.

Domain experts and data scientists experience the world in different ways. Domain experts operate in the business domain. Their habitat is a space made up of real business transactions and interactions, and a large part of the knowledge they rely on has accumulated organically in the form of expertise and experience, which they apply in tacit ways. To uncover hidden business opportunities, data scientists must capture the business domain in the model domain in the form of concepts, models, measures,
and hypotheses that are checked for their fit with available data. Any insight uncovered in this model domain then must find its way back into the hands of the domain experts to be put to good use.

**Benefits-Realization Process**

Success in realizing the benefits from data science will require a process that connects the business and model domains. However, this process isn't just a matter of collecting, sharing, or making available data and information— it's also about social connectivity. The latter has been underappreciated. Framed by the separation between the business and the model domain, I propose a benefits-realization process comprising the following activities: modeling the business, data discovery, operationalizing insight, and cultivating knowledge.

Modeling aims to represent a business idea regarding how to use data to improve the business in the model domain. It allows you to reason about the idea and interpret the available data. Discovery takes place in the model domain and is defined as a search for insight by running data experiments. It involves iterative data-based model analysis and synthesis. Eventually, the insights gained in the model domain are operationalized — that is, transferred from the model domain into the business domain and embedded into the work systems of your organization. Back in the business domain, cultivation promotes best practices for using data and a culture of decision-making based on data and analytics to maximize the return on your data science investment (see Figure 1).

![Figure 1. The data science benefits-realization process, which involves modeling the business, data discovery, operationalizing insight, and cultivating knowledge.](image)

This benefits-realization process should form the skeleton of all your data science projects. The four activities have one feature in common: they're guided by constructive conversations between the data scientists and domain experts. These conversations are characterized by interpersonal interaction and dialogue, letting project participants appreciate the perspectives of "the other side." Although these activities are equally
important, the emphasis changes throughout the project. Initially, more time is spent on modeling and discovery; later, more time is spent on operationalizing and cultivation. In principle, however, all activities are performed simultaneously throughout the project.

**Business Modeling to Create Common Understanding**

There's more to data science projects than haphazardly running data experiments and hoping for the best. The importance of modeling, especially during the initial step, can't be overestimated. Modeling helps data scientists understand the business idea's intent so the project stays focused. However, modeling shouldn't be treated as an ivory tower exercise. It should be a constructive alternative to "throwing ideas over the wall." So how can you make sure modeling adds value?

First, domain experts and data scientists should work closely together on the model. The business ideas that will be fed to the data scientists are likely to be diverse. Some might be wild, others more carefully thought through.

Some might be linked to a specific task, others might be more general. Some might address a problem, others an opportunity. Most ideas, however, will be loosely formulated — back-of-the-envelope style. Without any further clarification of a particular idea and its business context, data scientists are likely to misinterpret (and thus misrepresent) the idea, leading to the data science exercise's failure.

Modeling should be done in "conversation mode," so the data scientists can access the intimate knowledge held by domain experts — knowledge about the business ideas and their context. Any assumptions underlying the business ideas or bias in the data can be uncovered and challenged. Constructive, focused conversation is an effective way to gain a comprehensive, holistic view of the business context. One of the most important challenges data scientists face is trying to capture enough prior business knowledge in the model domain such that the insights produced by their data analysis goes beyond what domain experts already know or presume to know. Although the advantages are obvious, modeling in conversation mode requires some skilled handholding by the data scientist. Also, the domain experts must reciprocate and invest their time.

Furthermore, data scientists must be aware of the potential sources of bias. Suppose you want to predict the fraudulent nature of insurance claims by searching for fraud patterns in claims data. If you're relying on claims that have been labeled "potentially fraudulent" by claims handlers, then you should be cautious, asking exactly how the label was allocated. Your domain experts (in this case, the claim handlers) might have a biased view of the concept of fraud. It might be the result of a simple tick-the-box procedure they've been using to identify fraud. By talking to the domain experts and challenging them, the data scientists might uncover this bias, producing insights beyond what's already known.

In the same vein, asking your IT department to simply give you the data and leave you alone might not be wise, because, like your domain experts, the IT staff can be instrumental in pointing out any data quality issues and finding the right data.
The exploratory nature of data science projects makes managing the goals, scope, and expectations particularly challenging. There are definite risks of scope creep, even at an early stage. Although it’s good practice at the start of a data science project to spend some time considering alternative experimental setups, different possible models and data sources, don’t overdo it. You might be in exploration mode, but you’re also on a timeline to deliver benefits. Having domain experts and data scientists around the table, working on the model together, not only creates a common understanding of the business rationale but also secures the group’s commitment to a certain course of action and lets you retain control of the project.

**Data Discovery to Progressively Develop Insight**

In my model, discovery is a process of gaining insight using a technology-enabled sequence of fast data-based model analysis — that is, an algorithmic data search and output evaluation, using the model lens, and model synthesis, revisiting the model based on the results of model analysis. Each iteration represents a data experiment consistent with the scientific method. Conversations between your data scientists and domain experts direct the trajectory of this process.

Advances in computer processing, memory technologies, programming, and algorithms make it feasible to use big data. However, these technologies don’t just work at the push of a button. There’s likely to be as much exploration involved in figuring out how to make the technologies work (and work together) as there is in effectively pulling insights from data. Moreover, despite the impressive automation involved, discovering working business insights from data remains a human-centric activity.

The relationship between technology and humans during discovery resembles what Toyota dubbed “autonomation” — that is, automation with a human touch. In the Toyota Production and Lean Manufacturing system, the machine is stopped when an abnormal situation occurs, and the process relies on human intervention to solve the issue. Similarly, whenever a data search comes up with a potentially interesting insight, the machine should stop crunching and request human input. Guided by data scientists, the domain experts should assess and validate the output of each data experiment. What they learn should be the basis for the next iteration — adjust the model and then launch a new data search.

Consider, for example, the work my colleagues at KU Leuven and I carried out with the Amsterdam police department. Using text analysis of unstructured police reports, the police wanted to enhance their understanding of domestic violence. We were able to help them using not only an iterative process but also visually supportive data mining techniques and a versatile data navigation environment. The latter let us accommodate ad hoc data searches, so we could easily click through to individual data points from discovered patterns and seamlessly switch between aggregate and disaggregate data views. This, and the intuitive visual displays, made the conversations between data scientists and domain experts more productive.

Organizing for the incremental development of insight by interlacing consecutive data
Our solution to your specific needs experiments with synchronizing conversations between the model and business domains lets you carefully track the goals, scope, and expectations. It can also help you maintain the project's pace without being paralyzed by the sheer number of analysis options. This risk of "analysis paralysis" is countered or mitigated by time-boxing the discovery iterations and conditioning the termination of the discovery activity on the assessment of each iteration's marginal added value.

**Operationalizing Insight to Ensure Proper Action**

Operationalizing means equipping business operations and their management support with the right information at the right time and place. Many people prefer the term "business analytics" over "data science," because the latter carries the connotation of a lack of appreciation for the practical challenges involved. For them, the term data science refers to a situation in which the output of discovery is simply "thrown back over the wall" without proper attention paid to its use. Their aversion isn’t unfounded. No matter how powerful the insight discovered, if it’s not properly embedded in your work systems, then your investment will amount to nothing.

We can turn insight into action by adapting business processes, reorganizing teamwork, and improving decision making. In most cases, it will be the combination that proves most powerful. For example, in 2009, ING Belgium's marketing department set the standard for innovation with analytics by introducing a "fast-track campaign process." This happened in the context of a strategic effort to challenge the retail banking market by competing on analytics. The fast-track process operationalized ING's online marketing campaign experiments, and introduced "a business of experimentation" by providing a rigorous framework for selecting and testing campaign ideas and acquiring more knowledge about campaigning, customer profitability, buying behavior, and loyalty. Data scientists, together with communication and campaign execution staff, directed this test-and-learn process with a decision-making process that aimed to balance art and science. Note that operationalizing should be on the agenda from the outset. Conversations between domain experts and data scientists should ensure modeling and discovery are conducted with operationalization in mind.

Incidentally, it’s surprising that data science projects and business process management (BPM) projects are mostly conducted independently by different teams — BPM being the practical reference discipline for model-based development and improvement of an enterprise's work systems. Like data science, it uses tools and techniques to venture into the model domain to improve or reinvent the business. Modern BPM methods make a point of investing in process measurement, thus creating ample opportunities for data science. And while some approaches to BPM (such as six sigma and process mining) were born out of the systematic use of data and analytics for business process enhancement, many of these opportunities still remain untapped. It seems only logical to join efforts.

There is, however, another reason for data science projects to join forces with BPM teams. The field of BPM has had over two decades to learn what it really takes to transform a process model into a working process. BPM takes a holistic approach to managing change. It also comes with a real appreciation for boundary-spanning
project teamwork. Data scientists struggling to put model domain insights to work can thus benefit tremendously from the expertise and experience of the BPM team in successful operationalization.

An important caveat is in order, though. When combining data science and BPM efforts, make sure team members previously working on BPM projects adapt their way of working and managing to the nature of data science projects. The same holds for any IT people involved. The approach should support iterative, evolutionary development of insights, emphasizing collaborative model and business domain steering of the discovery trajectory. Don't allow the use of incompatible methods.

**Cultivating Knowledge**

It's critical to spread the word as much as possible with every data science project. Share the discovered insights, how they're used, and what you've learned. Extending and facilitating the conversations beyond project boundaries will promote an organizational culture of decision-making based on data and analytics.

The conversations during data science projects let the team members internalize the discovery experience. By engaging domain experts throughout the entire project, explaining and setting expectations by showing value in use and continuously articulating the learning, they can become data science advocates. Management can be expected to provide organizational support and proper incentives. Otherwise, project-based costing and a preoccupation with operational work will likely make employees dismiss cultivation as an "overhead cost."

That's why KLM Royal Dutch Airlines, in 2003, created an "ambassadors program" to support its customer relationship management (CRM) strategy. To reach out to each sales and operational unit in KLM, a network of committed and appropriately trained CRM ambassadors was progressively put in place to help coordinate local CRM developments and promote and facilitate CRM advancement across the organization. The program was designed to bring CRM close to KLM's employees, aiming for a viral approach to keep it cost-effective.

At SWIFT, the global provider of financial messaging services, big data and analytics are key areas for innovation. Back in 2007, SWIFT's innovation team started out as an internal software development shop for fast prototyping. Over time, driven by SWIFT's need for both product and cultural change, the team evolved into Innotribe, a "tribe of innovators" dedicated to promoting and facilitating collaborative creativity and open innovation in financial services. Described by the CEO as "the strongest brand SWIFT has produced in the last 30 years," Innotribe offers support in the form of education, perspective sharing, collaboration, facilitation, and incubation services.

Just like at KLM, successful stimulation of innovation has come to rely on ambassadors embedded in the different SWIFT departments. Carefully selected and clearly mandated, these internal agents spend about 15 percent of their time on innovation-related activities. Special attention is paid to communicate success stories and lessons learned. A variety of ceremonies, brown bag sessions, hackathons, and special staff events are
organized to expose as many employees as possible to innovation. Innotribe has its own social media strategy and works with a PR agency for external communication.

C-level executives have a special role to play in the cultivation process. In addition to advocating data science, they should also help create an environment in which effective "collisions" of ideas can happen. Much like product formation can only take place if there are effective collisions between reactant molecules, the generation of new ideas, knowledge, and insight requires the effective collision of ideas.

C-level executives should ensure the current organizational structure doesn’t block such collisions. At ING Belgium, for example, the marketing department was put in pole position to lead the way for the bank to compete on analytics. The Chief Marketing Officer started by putting his own house in order. To ensure the organizational structure matched the bank’s strategic ambition, he put the operational side of marketing communication and campaign execution on a par with the analytical side of campaign design, optimization, and evaluation, and he had business processes and teamwork redesigned to facilitate close day-to-day collaboration. At the same time, data governance — emphasizing the need for coordinating and stewarding the use of enterprise data assets — was put in place to stimulate the desirable use of data cross-departmentally in the organization. These and other organizational changes were used to infuse evidence-based decision-making and the use of analytics into the fabric of the organization.

The Platform

Returning to the interview with Lamba, he mentioned that Walmart is building its "Social Genome Platform to drive unexpected insights — and close the gap between decision makers and data scientists." What’s intriguing isn’t the fact that Walmart is building an analytics platform to help answer big business questions but that Lamba seems to expect the platform to break down the walls between data scientists and domain experts. The interview, however, gives little insight into how the platform will produce value from the data being amassed and scrutinized. Let me suggest some of the required capabilities, based on the benefits-realization process I presented.

Nurture Versatile Employees

The basic tenet underlying my view of benefits realization is that data scientists aren’t super heroes with all the qualities and knowledge to make a data science project successful. They must engage in many conversations with other parties to realize business benefits. The platform should be designed to support these conversations and stimulate the versatility of both domain experts and data scientists. They’ll probably continue to have their own habitat — either the model domain or the knowledge domain — but with a proper platform, they’ll feel more confident and at ease to engage in the necessary boundary spanning conversations. By facilitating this process of engaging, the platform should also serve to create a common language.
Produce Working Solutions

Data science exercises aren't about just publishing insights; they're about producing working solutions in the business domain. I've presented benefits realization from data science as a process composed of modeling, discovery, operationalizing, and cultivation. The platform should provide support for each of these activities and make sure that for every data science project, this process is followed and completed.

Support Information and Conversation Logistics

The traditional view on the IT platform supporting business intelligence — including analytics — emphasizes streamlining the information logistics chain — that is, the technical process of collecting and transforming data into information and intelligence. Traditionally, little, if any, attention has been paid to supporting the conversational side of benefits realization. The traditional view thus must be revisited. What we're looking for is a combined, synergistic support for information and conversation management. Ideally, the platform lets the model and business domains evolve and be enriched synchronously.

Foster Incremental Learning

Your analytics platform should let you develop an organizational analytics capability over time. Every data science project should be able to capitalize on the insights and lessons learned from previous projects and serve as a springboard for the next ones. Your data and model domain assets, as well as your business domain expertise and experience, should be readily accessible and reusable. Securing the long-term value of the platform requires involving enterprise and other architects in the design and use of the platform assets. Furthermore, conversations enabled by the platform will need to be governed to stimulate desirable use of the platform.

Create an Analytics Ecosystem

In the end, your analytics platform should form the basis for a digital business ecosystem. It should stimulate business productivity by opening up the world of data and analytics to the entire business community, and making it easy for internal and external parties to connect and innovate with data. It should offer robust support for running value-adding data science projects and making sure that every project contributes to collective learning. The platform, whose attractiveness lays not least in the continuous incorporation of new data sources and data science technologies, should also be an open invitation for ecosystem participants to venture into new domains with analytics.
Acknowledgments

This article was made possible with the support of the Deloitte Chair of “Bringing IT to Board Level” at Vlerick Business School.

References


Stijn Viaene is a full professor at Vlerick Business School and KU Leuven. Contact him at stijn.viaene@vlerick.com.
Our solution to your specific needs