Jul 29, 2015, 01:12pm EDT

What IT Needs To Know About The Data Mining Process

Meta S. Brown Contributor ⓘ
Tech
I decode analytics for business people.

Follow

This article is more than 6 years old.

No business can be data-driven if the only people interested in data analysis are the analysts. Just as the guidance of accountants and attorneys shapes everyday business, analytics must be integrated throughout the organization to provide value.

But when it comes to getting everyone on board, accountants and attorneys have a great advantage over data analysts. Financial and legal guidelines are supported by the law, contracts, and other strict rules, and it's understood that serious consequences may be in store for those who flaunt those rules. Analyzing data and using the results to guide business action is merely a good idea.

I've often presented talks on analytics fundamentals for IT professionals, and have found that these talks consistently draw large and engaged audiences. **Many IT pros are sincerely interested in supporting analytics, but don't know how to differentiate legitimate business needs from nice-to-haves, and are frustrated that data analysts don't understand and appreciate**

the reasons behind IT practices.

In "What Big Data Analytics Professionals Want From IT" I explained the most fundamental things that analysts need to do their job, things like raw (not aggregated) data and adequate computing power. Now, I'll take this theme a little deeper to show you the major elements of a good analytics process, and explain some ways that IT can and should be involved.

The Cross-Industry Standard Process for Data Mining, better known as **CRISP-DM**, has been around for more than a decade, and it's by far the most widely-used analytics process standard. It's an open standard, which anyone may use, developed by a consortium of over 200 interested organizations, with funding from the European Union. While it was developed specifically for data mining, it is flexible enough to suit many analytic styles. This process model is now so popular that diagrams from the original CRISP-DM documentation often crop up in presentations, without any reference to the original source of the material, or even to data mining.

The **CRISP-DM process model** has six major phases:

- **Business Understanding:** Get a clear understanding of the problem you're out to solve, how it impacts your organization, and your goals for addressing it
- **Data Understanding:** Inspect, describe and evaluate the available data.
- **Data Preparation:** Take data from the state it's in to the state needed for analysis.
- **Modeling:** Use mathematical techniques to make models

(equations or other logic) you can use to support business decisions.

- **Evaluation:** Figure out whether your models are any good.
- **Deployment:** Integrate models into everyday business.

It's not a linear process that starts with one phase and works neatly through each step in strict order. **These phases are parts of an ongoing cycle of analytics activity, and the analytics team may work back and forth among these phases frequently.** Roughly speaking, though, the process starts with a specific business problem and leads to creation of models and integration of those models into routine business operations.

IT has a role to play in each of the phases, though that role is much bigger in some than others. Getting access to the most relevant data in the most appropriate form clearly calls for involvement from data owners and gatekeepers. Integrating models into operations is almost always out of bounds for analysts; they must work with IT to make those changes. More subtle IT input happens in other phases, like modeling. IT staff generally don't have the skills needed to develop mathematical models, but they may have important expertise to offer about what business process changes are, or are not, feasible. That information provides analysts with a framework that they can use to determine what kinds of models may or may not realistically be deployed in the business.

Although many and varied data analysts use CRISP-DM, they **don't always understand and carry out all the elements of the process** in as much depth as they should. Each phase of CRISP-DM calls for several specific tasks to be executed and documented, but often people who say they use it skip over, or fail to properly document, some of those tasks. Even when the right work is done and documented, analysts don't always have good resources for managing the intellectual property they create.

Getting familiar with the CRISP-DM standard is valuable to IT professionals in a variety of ways. For example, one question that often comes up when IT hears about analytics data and computing requirements is: where's the business case for this? CRISP-DM requires identification and documentation of business issues, so using it (or an alternative well-defined process, such as the proprietary SAS SEMMA standard) ensures that **everyone shares a clear framework that outlines the business case, goals, work plan and outputs**. You'll know what's going on from the start, and have structure and documentation needed to demonstrate that you are doing the right things for the right reasons.

A defined process model supports IT's quest for proper management of data and work product. At least, it does if the necessary tasks are completed and documented. If everyone agrees to use the process standard, you'll be able to point out exactly what's required. And IT can and should take an important, often badly neglected, role by stepping in to **provide systems and resources for proper management of outputs** from data files to models to post-mortem project notes. Analysts don't always appreciate this kind of management, but everyone loses when the work is incomplete, lost, or inscrutable.

Working together with a shared and defined process helps IT and data analysts to understand one another, share ideas and communicate effectively. CRISP-DM is an established, respected and freely available standard that's available to everyone and adaptable to a wide range of analytics programs. Get to know it, and you'll do more, and better, work with analytics.

Where to learn more:

Original CRISP-DM 1.0 Step-by-Step Data Mining Guide is free for download.

IBM's slightly modified CRISP-DM 1.0 Step-by-step data mining guide, also free.

The original CRISP-DM guide is 75 pages of small type. If you find it hard to wade through, there's a simplified explanation of it (25 pages, big type) in my book, *Data Mining for Dummies*. Many public libraries have a copy (if not, ask if they'll get one), so you can read that free, too.

The Society of Data Miners is a young professional organization led by some of the pioneers of the industry.

Please note: 1) At the time CRISP-DM was initially created, I was employed at one of the organizations that participated in developing it. 2) I'm a member of the Society of Data Miners.

Meta S. Brown

Follow

I'm author of Data Mining for Dummies, and creator of the Storytelling for Data Analysts and Storytelling for Tech workshops. My work focuses on two challenges: 1)... **Read More**