database TRENDS AND APPLICATIONS

# Making It Measurable—Justifying Investments in Data and Data Quality for AI and Machine Learning

Dec 30, 2019

Seth Earley

Many organizations are experimenting with AI programs, but most of them face a significant and seemingly intractable problem. Although proof-of-concept (POC) projects and minimum viable products (MVPs) may show value and demonstrate a potential capability, frequently, they are difficult to scale.

One major issue is the quality, completeness, and availability of production data. POCs and MVPs are done in sandboxes with curated and cleansed data that is frequently adapted by hand.

However, it can be difficult to build executive support for and justify the need to make investments in upstream data processes. Senior executives would often rather put their organizational and social capital against something that is sexier than “data quality” or “governance,” such as applying AI to address a problem or better serve customers. The problem is that you can’t deploy the sexy apps unless the data foundation is in place.

One organization trying to create a 360-degree view of its customers encountered the following impediments:

- Sales and marketing technologies were disconnected.
- Basic analytic processes were not fully leveraging available data.
- The company lacked a clear understanding of the full customer lifecycle.
- Its data governance maturity was rudimentary at best.
- Data was in inconsistent formats across the technology ecosystem.
- No data curation and quality metrics were being developed.
- Ownership of data sources was unclear.
- No mechanisms were in place to monitor or enforce compliance with standards.
- Many analytics projects were not coordinated and lacked consistent approaches.
- New sources and formats lacked a standardized approach for onboarding.
- The complex technology stack had many stakeholders and users whose interests were sometimes in conflict.

Because of these issues, data quality, completeness, and consistency suffered. People did not trust that the data was up-to-date or reliable. Multiple efforts were made to fix the data issues downstream, but usually after it had already been consumed by some applications.

***For more articles like this one, go to the 2020 Data Sourcebook***

Because of the complexity of the problem, which had numerous causes and contributors, no senior executive wanted to tackle it. Although important, the task was neither sexy nor fun and it was not fixable through shiny new tools. Not only was the challenge too great but the organizational structure did not allow clear ownership of the problem and its costs or the benefits of a solution. The problem was woven into numerous processes and applications that spanned departments and functional areas. As I often say, “There is no budget for the greater good.” Though solving this challenge would have had benefits across the enterprise, it required sponsorship and accountability at the most senior levels of the enterprise.

## *THE PROBLEM OF DATA ACCOUNTABILITY*

Data and data remediation efforts are frequently considered infrastructure that is part of the cost of doing business, rather than representing something that can provide clear, measurable ROI. Because these efforts are difficult to tie to a specific business outcome, funding to solve the problem is difficult to secure. Data is also considered “an IT problem,” with little accountability on the part of the business side. But many problems cannot be solved by technology. One potential source of poor-quality data, for example, is salespeople who do not enter complete and accurate information into the CRM system. Yet, this problem is the responsibility of the business, not the IT department.

Each business unit needs to own its data curation, management, and quality, but this is difficult to implement. People may feel that they have no control over this or may lack the knowledge, resources, and technical capabilities to address the issues. On top of that, many challenges arise at the intersection of multiple data streams—for example, when integrating customer signals (data representing their “digital body language”) from various tools along the customer engagement lifecycle.

In one manufacturing company that dealt with a distributor channel as well as some direct consumer sales, these difficulties were systemic and significantly impacted revenue in a measurable way. Marketing processes had been in flux, moving from old-school use of traditional media and messaging to increased use of digital approaches, tools, and channels. The marketing team was not able to stay abreast of all of the rapid changes, and therefore business units and divisions took the initiative to experiment with and deploy proof-of-value projects using new tools (including machine learning-based applications, predictive analytics approaches, and cognitive ?agents such as bots and virtual assistants).

***For more articles like this one, go to the 2020 Data Sourcebook***

The problem was that these tools were deployed with significant technical debt and without thoughtful integration with existing standards, processes, and technologies. Not enough data was being captured about customers or their interactions throughout the lifecycle, information was not being shared with owners of other stages of customer engagement, and the lead-to-sales handoff was inefficient and lacked visibility to meaningful success measures.

## *SOLVING DATA CHALLENGES*

Clearly, the data challenges associated with customer engagement can mean the difference between success and failure—leapfrogging over the competition versus losing ground to them, or catching the wave versus being crushed by it.

How do you solve the problem? By taking the following steps:

1. **Understand the chains of trust for your data**. Very simply, this means mapping data sources for critical applications and determining who touches, enhances, or interacts with that data before it is consumed. Data remediation, measurement of value, and ROI should always be in support of a specific process. Identify the highest value process (there are numerous ways to prioritize where to start) and locate the sources and owners of each piece of data or content that is being used by the downstream process.
2. **Model your customer journey at a higher fidelity.** High-fidelity journeys describe customer intent in data terms and the information needed to produce the intended experience. Your systems can then respond with the next best action and next best product or content. This process is the optimal way to determine what source systems are most important for each stage of the lifecycle.
3. **Map your chain of trust to each stage in the customer lifecycle.** Mapping data to the lifecycle means understanding what data is needed when, and what data your systems are using to inform the customer experience. You may find new sources or identify sources that are less important to a particular phase. You may also find that some data cannot be readily accessed, or that missing data is causing unnecessary friction at a point in the journey.
4. **Establish cascading metrics for each stage and process of the customer experience.** Cascading metrics start at the data quality level and then move on to process metrics that require the data. Metrics are then aligned with business outcomes, and the business outcomes aligned with the enterprise strategy. The important piece is to show the linkage of strategy, outcome, process, and data. Otherwise, it will be impossible to know how data initiatives impact revenue or costs.
5. **Tie the appropriate metrics to business outcomes and assign responsibility for those outcomes.** Identifying responsible parties should be straightforward. Every critical step of the customer lifecycle (acquisition, purchases, service, financials, loyalty)

should have an owner, and that owner should be measured on results. This step establishes responsibility for the data as well as outcomes. Connecting metrics to business outcomes shows how data supports or inhibits the results.

6. **Build a decision-making structure for vetting and approving changes at the correct level of granularity.** Governance is not a word people typically associate with responsiveness and adaptability, but it can be. The key is making the punishment (the meetings) fit the crime (the importance of the change) and to involve only those people who are in the chain of trust—and no more, no less. Since the outcome impacts them directly, they will attend and participate. It is also useful to get people to sign off on a responsibility matrix to ensure they understand their commitment and what is expected.
7. **Build a set of decision-making processes for changes.** Decision-making processes go along with structures. Several questions should be answered: Who is the primary decision maker for different categories of changes? How do you determine what is a major change and what is a minor change, and does the decision-making responsibility vary in each case? When do others have to be informed or involved? Who else needs to approve the change? When do changes need to be user-tested or regression-tested?
8. **Enforce compliance with processes.** Having standards without compliance is similar to having vision without execution. In both cases, the result is delusion. To make standards meaningful, be sure to address the following issues: How will an enforcement rule be carried out? What are the gating factors? Who should things escalate to? What is the reporting process? Don’t overlook the key step of educating employees about the standards and why compliance is so important in order to ensure buy-in.

## *WHAT’S AHEAD*

For data programs to show real value and not be considered a science experiment—or worse, a waste of scarce resources—the effort has to be tied to something measurable.

Building out these processes will ensure that the right resources are applied to data initiatives, which are a fundamental requirement for AI and machine learning. The processes will allow measurement of ROI and provide the supporting data for adequate resourcing.

Emerging technologies, including machine learning and AI, run on data. In fact, for these technologies, the data is more important than the algorithm since the algorithm will not work with poor data, no matter how much it is tuned.

Even unsupervised learning programs benefit from labeled data. Labeled data is architected data, and information architecture managed by metrics-driven decision making is critical. There is no AI without IA (information architecture). This is not magic: It all depends on having a strong data foundation right and linking that needed foundation to capabilities that yield measurable business value. Only then will your AI, advanced analytics, and digital transformation programs be successful.

http://www.dbta.com/BigDataQuarterly/Articles/Making-It-Measurable-Justifying-Investments-in-Data-and-Data-Quality-for-AI-and-Machine-Learning-135513.aspx http://www.dbta.com/BigDataQuarterly/Articles/Making-It-Measurable-Justifying-Investments-in-Data-and-Data-Quality-for-AI-and-Machine-Learning-135513.aspx?PageNum=2