EXPLORIUM

It's Time to Optimize Your Analytics — Why You Need a Data Acquisition Strategy

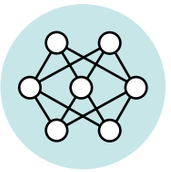
Table of contents

Solving a problem requires two things — having a solution in mind and the assets you need to implement it. Across industries and fields, we're discovering that for a large number of business and organizational problems, that asset is increasingly data. This isn't anything most of us haven't heard before, but it's taking on new meaning as we discover just how impactful having the right data can be.

Today, it can be the difference between an organization failing in various ways and creating meaningful insights that can drive ROI. More importantly, as analytics tools become commoditized, relevant, useful data is becoming the key way companies can differentiate themselves and find new market share and revenue sources. More than ever before, having the right data to answer the questions at hand is an indicator of whether an AI transformation or initiative will succeed — more so than which model they're choosing to use. As such, the companies that are adopting this new mentality are positioning themselves to succeed. More importantly, they're proactively seeking out new data sources the same way they would seek revenues and new customers.

The question, then, isn't how much to spend on a fancy new ML model or how to build more data science tools, but how to build a data acquisition strategy that ensures the rest of your initiatives run smoothly. Because it's not just about having more data — at this point, most companies already produce large troves of it every day. It's really about understanding what data you need, from where, and finding the most efficient ways to gather and leverage it.

This whitepaper will cover why advanced analytics are moving away from a focus exclusively on the tools we use for analytics and in favor of having the right data and automating the processes to acquire it. It will also cover why a data acquisition strategy is crucial for your business success. We'll also cover why using alternative data as a key element in your data strategy is a vital component of your analytics initiatives.



Models can still be differentiators — but only with the right data

Data science and analytics have quickly matured in the past few years. Ten years ago, having the right BI platform or a robust ML model could make a huge difference. Companies could build and fine-tune their models to get better predictions and find new revenue streams. Model optimization was a valuable strategy, and it produced great results — until it didn't. The problem with model optimization is that at some point, you're facing diminishing returns in terms of what you can keep squeezing out of your models. Put simply, there's only so much you can tweak before you're just changing things around with no real purpose.

On top of that, the AI ecosystem has matured to a point where the tools and models we use to make predictions have become commoditized. This is the result of a confluence of a couple of factors:

- **The transition from BI to AI:** One of the biggest leaps in analytics has been the proliferation and improvement of AI tools that permit predictive analytics as opposed to simply looking at historic data. The difference might seem semantic, but in reality, AI allows organizations to view their data in new ways and extract value. Working manually, it might take a team several years to go through an organization's data catalog to find relevant patterns. For an AI-powered ML model, the same process could take a few minutes and produce much broader and better results.
- **The shift from on-premise to cloud computing power:** ML — especially at enterprise levels — requires computing power, and on-premise infrastructure is a major roadblock for many organizations that would otherwise be adopters of the technology. However, with the increased access to cheap, powerful cloud computing, this is no longer really a problem. In fact, most organizations can quickly spin up thousands of servers with a click to match their computational needs.
- **The rise of open-source statistical tools, both academic and commercial:** The third major factor at play in how ML has become commoditized is the easy access almost anyone has to the tools necessary to build new models. The massive expansion of open-

source tools means that anyone — from a data science Ph.D. to a Python novice — can build powerful models and find deep insights with 10 lines of code and a freely available dataset.

What these three factors add up to is that building models today isn't where the real magic happens in data science. Indeed, it's safe to say that with the amount of tools and platforms available, it's hard to really set yourself apart from the competition simply by having a "better" model. Instead, organizations looking to find new sources of revenue, opportunities, and market share, need to look to the other side of the equation — the data they're using — to get an edge.

Today, being a data scientist is less about understanding each element of a statistical model or knowing every single algorithm possible and more about seeing the broader picture. A successful data science team doesn't need to have a deep knowledge of analytical modeling theory — they can rely on decades of shared expertise. Instead, a successful team will leverage creative thinking, ML models, and the right combination of alternative and internal data to gain better insights, find new avenues to expand, and discover new revenue streams.

Bottom line:

In an age where ML models and analytics tools are increasingly commoditized, the real key to unlocking an organization's AI and data science potential is having the right data to feed into your advanced analytics tools and platforms.

How do you find the right data? Building a data acquisition strategy

Buying data is a crucial step in optimizing your analytics, but it's also not a one-time step. Instead, data acquisition is a process that constantly occurs, and requires continuous oversight and consideration. So, if having the right data is so important to organizations' AI and data science success, it makes sense that everyone has a well-established data acquisition strategy, right? Well, not so much. The fact remains that while data is becoming significantly more important, the way most organizations and companies go about finding it remains, to say the least, suboptimal. However, this is not entirely the organizations' fault.

The solution, though, is not just to slap something together or to just buy more data. The right way to do it starts with understanding **why** you should have a data acquisition strategy in the first place.

It's not just about piling on data; it's about getting the right data for your objectives. Your data acquisition strategy needs to consider what problems you're trying to solve, whether it's attempting to solve an unexpected — or even "unsolvable" — problem or optimizing an existing mission-critical analytics solution:

- **New and unsolvable problems** need new data because you need a new perspective that's not available with the information you're already using, or simply because you don't have enough data to find an answer. Let's say your organization sells products to SMBs, and you want to predict which companies might go out of business in the next year to make sure you're targeting the right customers. In this case, your first-party data wouldn't be of much use in a predictive model — you don't have insights into organizations' financial health and revenues. On the other hand, business filings, financial reports, and transactional data could help you answer these questions more successfully.
- **Optimizing existing solutions** also requires you to look for data, but in a different way. Here you're not looking for a completely revolutionary breakthrough — in fact, most organizations would look at an uplift of even 1% in their models as a major success. A retailer

might already have a successful model to target its customers, but adding seasonal data, user demographics, weather information, and footfall traffic could help make these predictions better, leading to higher sales volumes, better revenues, and more engaged customers.

No matter what you're going to use the data for — to solve an "unsolvable problem", to optimize your model, or even to build a completely new model — there are two key facts you need to be aware of: what data can do for you, and how the data landscape is expanding.

Data is a competitive advantage

Imagine two online lenders competing in the same space, both using ML heavily to mitigate their risk. It's safe to say that even if there are some variances, their models will look very similar. However, one of them has a much better success rate when it comes to extending loans than the other. The difference isn't the model, but rather what they're training their models with. Why? Because having better data lets the organization gain a better perspective, provide more accurate features, and understand risk factors more reliably.

The key here is that, all things being equal, a company with better data will have a much higher success rate with AI and data science than an organization that focuses too heavily on building and optimizing their models to the exclusion of everything else. Better models might be slightly faster, but better data lets you make smarter decisions. In this case, data is the "alpha" — the differentiator — that organizations need. No matter if you're using the exact same model, if you have better data, you'll get much more out of them.

There is explosive growth in the volume and variety of external data sources

We all know that organizations create and collect troves of data today. What's less obvious is that the number of external and alternative sources has expanded at a much faster rate, and it's growing every day. And it's not just volume — the variety of data available today extends to data types, providers, and dimensions. For organizations that aren't really prepared for it, this can be overwhelming and make it easy to miss out on great data simply because they don't know where or what to look for. This might not seem like a huge deal at face value, but it could mean the loss of thousands of dollars in earning potential,

hundreds of new customers, or even the ability to operate successfully.

The question isn't how to get all the data you need now, but rather how to build processes and systems inside an organization that will make it easy to find the right data when it's needed and build it into the data science process. Organizations that are truly data-driven know that they can't afford to miss out on an opportunity simply because they didn't find the right dataset.

Therefore, to succeed in today's AI world, you need to develop tools, processes, and best practices that let your organization quickly search for, discover, integrate, and monitor the external data sources they need. This means crafting a well thought out, methodical, and systematic data acquisition strategy.

Unlocking the power of alternative data — the key components of a successful data acquisition strategy.

Building a successful data acquisition strategy that makes your organization a "data hunter" (that is, an organization that's constantly on the lookout for new data opportunities) is more than simply googling "data for X". It requires having the right people and teams in place, automating key processes, and finding the tools you need to help them succeed. Let's dive into what makes a good data acquisition strategy:

Streamlined data acquisition:

To optimize your pace of innovation and experimentation, you need to find the right data, and do so quickly. This means building established methodologies at several points:

Search processes:

Your organization should have an established role for a "data hunter," a data acquisition manager whose key function is to constantly look for new and valuable data sources for existing and new projects. This means your organization isn't reacting to new data needs, but preempting them, and constantly adding new data sources to be tested, measured, and potentially acquired.

Agile procurement:

Once you've found data, you'll need to procure it, which is different from traditional software processes. Your organization needs to clearly understand how the process works to cut down on the — unfortunately standard — long timelines in the purchasing process. This includes two key phases:

- Pre-decision testing, which can include everything from a simple commercial agreement to DPAs, NDAs, and other legal agreements that can take some time to negotiate before you can even acquire a testing sample.
- Buying or licensing, which involves finalizing terms, agreeing on

prices, completing the legal process, and finally acquiring the data you need.

Testing:

This is a key component and one that can create bottlenecks if your organization isn't careful. While some data vendors might offer truly unique datasets that require you to dive deep into them, many offer similar data types and sets, meaning you can likely create automated processes to test them. However, even when dealing with similar data, you might be faced with some more expensive data that might look better but could actually provide much less coverage than a cheaper, seemingly less valuable vendor.

Therefore, you need to have efficient, objective processes that cover three phases:

- **Coverage:** This is all about making sure the data or API you're acquiring matches with your data. For instance, if you're looking at demographic data, you need to ensure that it covers the regions and geographical areas you're analyzing.

- **Quality:** Data quality means how accurate values are inside the data, how many missing or null values a dataset has, and how up to date it is. For example, if a B2B dataset includes Apple but claims the company has 50 employees, you can likely discard the dataset as low quality.
- **Relevance:** Finally, even if the data has great coverage and accuracy, it still needs to be relevant to your objectives. Just because a dataset matches your data doesn't mean it'll help you do what you're aiming to achieve. Admittedly, this is a little more art than science. Let's say you're buying data for a commercial risk and underwriting ML model. If the data you're looking at only offers one attribute — the company's logo color — it might be both accurate and have great coverage, but it won't really give you much of an uplift.

Automation and scaling search and discovery

As your data needs scale, you'll need to have a greater variety of data, which leads to its own problems. You simply don't have time to test, analyze, and think about every dataset you could have for every project you're running. This leads teams to overlook existing data sources and opportunities. Instead, you need to adapt by finding scalable ways to carry

this process out. Today, this can be easily achieved by using data discovery platforms that can quickly scour thousands of data sources and provide you with the most relevant data without the difficulties that come with it. There are several advantages to using a single data platform that connects you to multiple data sources without having to integrate to each vendor separately:

Security and compliance

One of the biggest headaches attached to finding new data is ensuring that it can meet an organization's stringent security and compliance requirements. For example, many premium data sources come in the form of APIs that need inputs that could be sensitive (personally identifiable information [PII], for one). To enrich a dataset about individuals, you would need to share this information with your provider to match the data. Of course, there are agreements, contracts, and privacy policies that make the process safer, but when you need to do it with tens of organizations simultaneously, it becomes untenable.

The rate of data asset growth

The more data you need, the more agreements, negotiations, and procurement processes you'll have on your hands, and this isn't a good

thing. It makes it harder plan your budget, your data usage terms, and more. Moreover, the process is time-consuming and long, meaning you'll need to invest heavily into getting it done — for every single data source you use.

From an integration standpoint, each vendor has a unique schema for their data, API, delivery, and authentication methods. To handle every single one, you'd need to have an entire department handling all of this full time.

Instead, hiring a single platform that can take these processes off your hands and automate them without sacrificing quality and compliance can make it a much more feasible solution.

Facing specific data type challenges

The above points cover broad strategy concerns, but data acquisition also needs to focus on the specific data you're acquiring. Every dataset has unique requirements and challenges that you need to confront to successfully integrate and make the most out of it. For example, consumer data requires significantly more regulation, due diligence, and it focuses on data that is constantly shifting (PII). On the other hand, geospatial data is largely available open-source, but finding valuable datasets is much like finding a needle in a haystack.

The habits of enterprises that implement successful data acquisition strategies

Knowing what we just learned, where does that leave us? With the fact that organizations can't just hire someone to find data and let them go wild. Embracing data discovery as a key driver for success requires organizations to go all in and truly transform the way they approach ML and data science. First of all, it means focusing on three key areas:

People

The most basic — and important — question is "do we have the right people in place to expand our organization's data assets?" Your data acquisition strategy starts and succeeds with having a team that understands your organization's data needs, and proactively looks for opportunities to get better data and better results.

Process

This goes back to actually having a data acquisition strategy in place. It's not enough to know that you need data and then go look for it. You need to build a pipeline that starts with discovering data but includes testing, procuring, licensing, and integrating that data effectively and quickly to actually give you the uplift you're looking for. It also means understanding the challenges around data acquisition and successfully navigating them every time you need data.

Tools

Finally, you need to give your team the right tools and platforms to succeed. They can certainly deal with datasets and providers as they come, but this will only create bottlenecks. They need to have tools that let them quickly find the data, test it for coverage and relevance, and integrate it without spending weeks or months going over every detail. It's about unlocking the value of the data sources available. So what makes a successful data discovery platform? A few things:

- **Access:** The bottom line is that your chosen data platform should significantly expand your access to data sources. It should include not just a single source or a few, but hundreds or even thousands of data sources that can be instantly integrated into your ML models or AI projects.
- **Automated data discovery and feature engineering:** More importantly, you don't have time to scour through thousands of data sources. Your chosen platform should do the heavy lifting for you, finding the right data for your chosen use case and even extracting the most relevant features to give you the best insights.
- **Production pipelines:** Your data doesn't work in a vacuum, it needs to be integrated quickly and smoothly into your production-ready models at scale, and it needs to be monitored and adapted on the go. Therefore, your data platform should be able to build data pipelines that are reliable but malleable, and can scale according to each project's needs on-demand.
- **Auditing capabilities:** Crucially, you need to understand exactly how a data platform is performing, and what uplift it's actually giving you. A good platform will include tools to help you both understand how each data source will boost your existing datasets, and how

that in turn will give you better insights. It should be able to measure coverage, uplift, and the relevance of each data specifically to your predictive or analytic problem.

The future is data — are you prepared?

Data science is moving away from model optimization as the major way to differentiate yourself. Your organization can't afford to build data acquisition processes that are too lengthy or too ineffective. Even so, data acquisition and deployment takes time, and you simply must account for it when you're creating a strategy. How do you cut down on this time? How can you conceptualize processes that will work effectively without taking too long?

Developing the right methodologies, processes, and mentality requires you to consider how data works with your models and your goals. More importantly, you need to create a system that lets you constantly look for better data and acquire it before it becomes an emergency. By finding the right centralized, integrated platforms and having the right processes in place, you can quickly give your organization the uplift it needs, and ensure your ML models and AI tools will give you the results you're looking for.

About Explorium

Explorium offers a first of its kind data science platform powered by augmented data discovery and feature engineering. By automatically connecting to thousands of external data sources and leveraging machine learning to distill the most impactful signals, the Explorium platform empowers data scientists and business leaders to drive decision-making by eliminating the barrier to acquire the right data and enabling superior predictive power.

For more information, visit www.explorium.ai