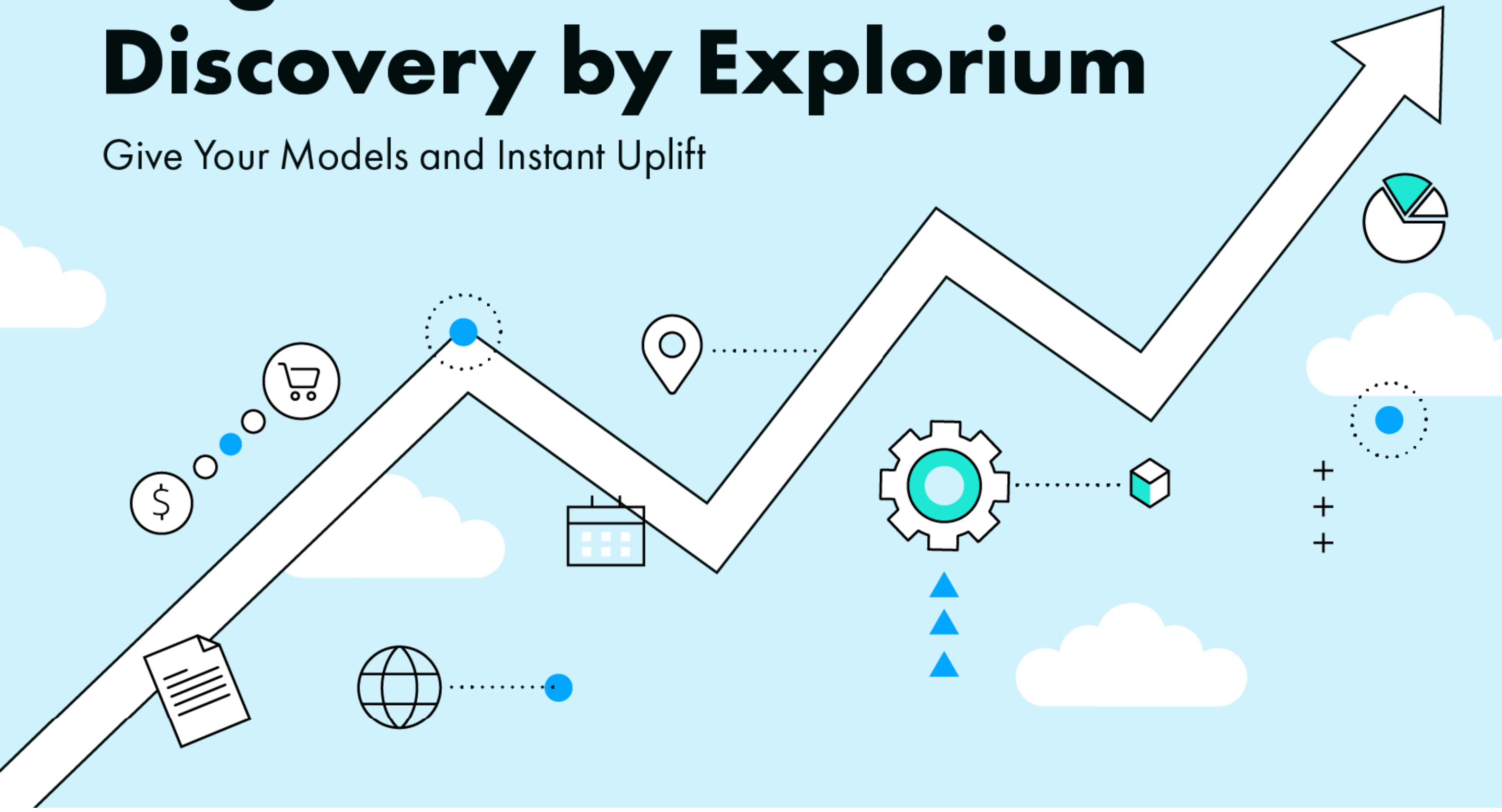
Augmented Data Discovery by Explorium

Give Your Models and Instant Uplift



Augmented data discovery with Explorium gives you access to thousands of external data sources with a single click.

Machine learning (ML) models thrive on data. Use good data, and they'll produce great results; use bad data, and well, you know what comes next. This primacy of data means that you need to pay extra close attention to the start of your ML lifecycle. Long before you start building your algorithms, you need to make sure that the data you'll use to train and test them will be enough to give you the best results. However, the question is whether the data you already have — your internally collected dataset — is actually the data you need and if it's enough to give you the insights you're looking for.

The problem: finding the external data you need, when you need it

It's not that the right data is nowhere to be found. Quite the opposite, actually. It's easy to argue that there might even be too much data out there for a single organization to sift through. For organizations looking to deploy ML models and data science at scale quickly, going it alone means having to dig through an infinite number of haystacks to find a couple of needles that may or may not be relevant and provide an uplift to their models.

The key problem, aside from the sheer amount of data to look through, is that finding the right sources is a costly process in terms of money, time, and resources. Moreover, the complex data market means that data quality is hard to determine when searching through thousands of possibilities looking for the few data points you actually need.

A Gartner report found that among the largest roadblocks to data acquisition and discovery, decision-makers noted data quality, a lack of data standardization, and simply sifting through a massive catalog. A separate report found that **only 8% of its respondents claimed to take less than one month integrating their external data, while 31% of responders claimed the process took over three months. With over 81% of companies surveyed noting that integration was a challenge, finding ways to resolve them is critical.**

What tools exist to fix these problems today?

While there are serious challenges to getting the right data, there are some tools available to organizations, even if they don't resolve all their problems. However, the available solutions are piecemeal, fixing smaller symptoms without resolving the core issue — the cost and time investments required to discover, test, and integrate external data quickly and scalably.

There are two existing solutions out there, both with their own issues. The first is the data catalog, in which private providers collect massive numbers of datasets and give users access to all of them but don't offer the tools to integrate it. The other is the data integration platform, which gives you the tools to connect your external data to your internal datasets, but doesn't help you find the data you need. This is where Explorium comes in.

	Data catalog	Data prep tools	Explorium
Access to data	✓	✗	✓
Data integration	✗	✓	✓

Augmented data discovery with Explorium — connecting the dots

The goal of using a platform for data discovery and acquisition is to significantly cut down on both sides of the equation — getting the data, and finding the most relevant points to connect to your internal sets. Explorium takes a holistic approach to data discovery, giving you both immediate access to thousands of proprietary, premium, and public data sources while also instantly connecting them to your datasets. The result? Faster, better, and more efficient data discovery that supercharges your machine learning initiatives.

Everything for your external data needs in one platform

Explorium's augmented data discovery starts with your data, but by the time it's done and you move on to enrichment and feature engineering, you'll be dealing with a completely new — and enhanced — dataset. The secret? A powerful engine that gives you the right data on demand. Let's look at an example to see how Explorium works.

Say you're the CIO at a retailer, and you're looking to get some new insights on your customers to predict changes in the market better. You collect data from various touchpoints. This might include things like number of page views, items in shoppers' carts, the number of clicks prior to a purchase, and the unique pages and items users have viewed. You'll likely also know their purchase history, which can inform their preferences.

Your analytics are okay, but they're not great at making reliable predictions, and they sometimes show biases that can impact the insights you're receiving. Let's imagine you're building a model to predict repeat buyers during a promotion, but it's harder to do when you're dealing with customers you've never seen before. Using Explorium, you can quickly add new data to add context and make your model more powerful in a few minutes.

- **The first step is understanding your dataset.** Once you upload your dataset, our platform will explore it in-depth to fully understand what you have, what you're missing, and what you need. The platform uses a variety of exploration and matching methods to audit your data to start building the right ontologies — the data types of each column in your database — clean up any null values and redundancies, and other small factors that could impact your data enrichment in a meaningful way.

Think about it this way. Your retail store has multiple touchpoints, both physical and digital. You collect a lot of the same data on both fronts, but how the data is input, how it's organized, and even what's been left out will change. Your store manager might input names as [last name, first name], but your CRM might collect it as [first initial, last name]. It might seem small, but without resolving it and standardizing your data, your enrichments won't be as impactful. Explorium handles this as soon as you start the discovery process.

- **The second step is to scour the Explorium Enrichment Catalog.** Don't worry, though, you don't have to do it manually (actually, your job is done once you press "start", all that's left is to see the results in a few minutes). Based on your existing data and columns, our ML algorithms will scan through thousands of datasets and test them to determine which is most relevant to your core data and your predictive question. Remember, it's not always about simply having more data — it's about having the right data to get you results.

  Let's return to your hypothetical retailer for a second. When you first run the data pipeline on Explorium, the platform will instantly connect you to hundreds of data sources. For instance, Explorium might connect your data to:

  - **Social media data** that could add context about users' preferences and habits, as well as the interactions and engagements they've and different brands. Additionally, aggregate social media data is a great way to understand the conversations your target market is having, giving you better tools to predict trends and understand who might be more amenable to become a repeat buyer.
  - **Weather, event, and news data** that could highlight potential times when buyers might become repeat customers due to external conditions — a major social movement or viral trend, a snowstorm that limits access to brick-and-mortar stores, and more. This can be layered with other data types to create powerful predictors that could give you better insights.
  - **Socioeconomic data based on region and ZIP code**, which includes things like average home rental prices, average mortgage rates, and average household income. These can generate predictions based on where a user is shopping from, and can help you target specific campaigns and ads at users who are more likely to make multiple purchases based on their socioeconomic status.

- **Now, it's time to enrich your data with the external sources Explorium found.** Here's where the magic happens, and now you're ready to finally get to the next step in your data journey. Once Explorium has sorted through, ranked, and selected the best data sources for your predictive questions, it's time to combine your datasets with ours. The good part is that your data has already been prepped and matched to the right data points in the external sources you'll be using. Additionally, the platform will automatically scan both your data and ours to ensure no redundancies, overlap in columns and rows, and no missing values. Most importantly, the platform scores all the data discovered based on coverage and its relevance to your use case.

  For the last time, let's go to your imaginary retailer. By now, you can see hundreds of potential sources (let's put a number on it, and say roughly 500 possible matches). Now, remember, this doesn't mean all 500 are actually relevant or even impactful. Instead of overloading your dataset with more, the enrichment process cuts down those 500 sources into the 20 or 30 that actually give your existing ML models an uplift and thus better insights. Now, you're not dealing with 31 sources, but rather a single enriched dataset that's ready for feature engineering and later training and testing.

Sounds like a lot, right? That's because it is, but unlike doing it manually, this will only take you a few minutes of waiting rather than weeks or months of headaches. Don't take our word for it, though. Let's see why Explorium is better when you need to find the right data.

The Explorium effect on your data science

If you still don't believe us, let's put it this way. You can definitely do all of this on your own. You can have your data science team scour the web and everywhere in between to find the right data for your machine learning models. And they'll probably succeed — eventually. However, getting there will take a significant amount of resources, and you may miss out on some nuggets of gold due to the sheer difficulty of exploring so much data. Here are some of the biggest ways Explorium supercharges your data science:

- **Cut down on grunt work, and focus on the tasks that matter.** It's not that having the right data isn't important — quite the opposite. But the process of getting that data manually is time and resource-intensive, and it can impact the rest of your ML development. A study found that data scientists spend nearly 80% of their time on data preparation — collecting the right data and molding it into something functional. This means they have less time to conceptualize, optimize, and generally build the best possible ML models to answer your predictive questions. Automating your data discovery frees your data science teams to devote their full attention to giving you the impact and ROI your organization is looking for.

- **Guarantee you'll get the most relevant data for the most accurate predictions.** Your ML models are only as good as the impact they have on your business. Therefore, they deserve the proper amount of attention (and data). It's not about if you need external data; it's about understanding what data you need and how to get it. Going it alone means you'll have to limit your search to what you can vet on your own. This means you'll be missing out on hundreds of data sources that could give you an unexpected perspective and a major uplift in your predictive capabilities. Explorium puts thousands of diverse data sources at your disposal and does the hard work of vetting, cleaning, and preparing the data to give you an instant impact.

- **Build scalable and up-to-date ML pipelines.** Your business landscape isn't static, and your ML models shouldn't be either. Augmented data discovery with Explorium lets you build dynamic models connected to the most up-to-date and relevant data for your organization's predictive questions. More importantly, keep your data pipelines ready to scale with access to thousands of continuously updating and expanding external data sources. Stop wondering what you could have done with more data, and find out in minutes. More importantly, keep all your data management under one easy-to-access roof, avoiding the need to constantly deal with data providers when you require new or different data.

Give your models the data they deserve

If you want the best results from your data science efforts, you need the right tools for the task. Explorium's data science platform instantly connects you to thousands of pre-vetted and ready-to-use external data sources that can make a real impact on your ML models.

However, more than just access to external sources, Explorium's augmented data discovery gives you exactly the data your models need and automatically connects the most relevant data points. This way, you can cut your prep time from months to minutes, and focus on the real labor of building the best possible ML models to give your organization a great ROI.