EXPLORIUM

Rethinking Data Acquisition with Explorium

Table of contents

The key to solving any analytical problem is having the right data.

Data is an asset. One that forward-thinking organizations seek out just as actively as they would revenue streams or new customers.

And for good reason. With the right data, organizations make smarter decisions and solve key business challenges. Today, “good analytics” means going well beyond a few commonly available algorithms or a flash dashboard to showcase internal data.

Having the right, most relevant data is now the key success factor in an organization’s AI and analytics initiatives.

This also means that the term “AI arms race” is really a misnomer. What we’re living through is a data arms race. Enterprises looking to make the most out of the intelligence revolution must evolve. They must build effective strategies to find, acquire, leverage, and understand new alternative data assets.

The logic is simple. As data becomes increasingly central — the “new oil”, so to speak — and as the tools to analyze it and refine it become more commoditized, the best strategy is the one that offers the widest access to valuable data sources.

With a data acquisition strategy, organizations can search, acquire and leverage more relevant data assets. This lets them solve complex problems much faster. At last, they become truly data-driven.

It’s also why data-fuelled enterprises increasingly capture market share.

Being data- and AI-driven doesn’t just tell organizations what’s happening right now., It also helps them figure out what’s likely to happen next, so that they can prepare, anticipate problems and seize opportunities as they emerge.

In short, the buzz around AI is justified.

You simply can't understate the impact this technology offers companies that leverage applied AI techniques for analysis and predictive modeling. Across all kinds of industries, those that embrace AI and ML are solving more problems faster, growing their business, winning new customers and developing lucrative new revenue streams.

But without the right technology - and the right data - many projects never get off the ground.

The harsh reality is that most AI initiatives fail long before they ever reach the algorithm stage. That said, the sheer breadth of open source tools, along with innovations in cloud technology, have simplified predictive algorithm design - once a major limitation. This forces companies looking for an edge to go beyond their models and tools, identifying new areas for improvement.

The biggest of these is data. Organizations must now figure out how to operationalize their ability to access and leverage the right data for any business challenge.
In this whitepaper, we'll reveal the new paradigm emerging in data asset strategy. We'll also present solutions for data teams looking to adopt the model across the entire alternative data acquisition cycle.

Most importantly, we'll tackle one of the biggest challenges for enterprises this decade: finding the right data to meet your business objectives.

How technology shifts are driving data's value

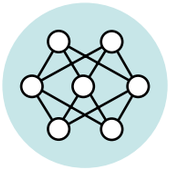
Automated data discovery is a new approach to an old analytics problem. You take a question or model-in-design and ask: what's the right data to select and use? The quest for data boils down into three core ideas: understanding why data has become important, realizing the power of alternative data, and building a data acquisition strategy that can successfully leverage that power.

In this whitepaper we'll focus on:

- **The shifting paradigms in data science and analytics.** As AI and analytics tools become commoditized and democratized, organizations must focus on the "data" side of data science to differentiate themselves and get ahead.

- **The need to have a clear data acquisition strategy.** Instead of continuing to purchase data ad hoc, enterprises are rushing to build more defined strategies and processes to acquire data. That includes finding the right platforms to discover it. This stems both from the newfound value of data and the rapid expansion in availability of external and alternative data assets.

- **How to leverage the power of alternative data.** This includes the components of a successful data strategy, the production challenges that you might face, and how data helps you solve different kinds of challenges.



The Emergence of Data as the Differentiator

Data in the face of analytics tool commoditization

More than ever, data is the factor that decides which enterprises will claim the largest market share.

The data revolution is deeply rooted in a decade's worth of transformational trends that democratized and commoditized analytics and BI:

- **The new balance between BI and AI.** Business intelligence (BI) remains a key component of organizations' analytics activities - and for good reason. It offers great visibility and visualization, along with flexible reporting. However, AI and data science have shifted decision-making power towards AI tools, which let organizations discover patterns and organize data in a matter of minutes.

- **The transition from on-premise to cloud compute power.** One of the major factors driving AI's growth is the ease of access to computing power and cloud elasticity. Public cloud infrastructure providers let companies quickly scale up their available resources on-demand. This makes it easier for data scientists to support innovation.

- **The emergence of open source statistical tools.** The expansion of open source has brought decades of research and years of expertise into everyone's reach. Today, anyone can write a ten-line Python script which could easily train a predictive model with millions of freely available data points.

Combined, these trends add up to a generational, transformative shift for the field of analytics.

Formerly complex fields that were once the exclusive domain of the few, such as predictive modeling, are now commonly accessible. AI, computational power, and reporting tools have been commoditized. All of which creates a massive opportunity to solve an ever-growing set of business challenges quickly - and scalably.

No wonder more and more business functions are leveraging predictive algorithms and advanced analytics. No wonder they're using this to automate and optimize a staggering variety of business problems. The barrier to creating powerful analytics-driven solutions is now lower than ever before.

These days, it's rare to find a data scientist building a new algorithm from scratch. The idea of spending time on "negative" engineering — doing tasks that can be easily solved or already have working solutions — is outdated.

Instead, data scientists spend their time implementing and testing a variety of easily accessible algorithms and equations. Instead of focusing on the processes, data science teams are focused on tailoring existing systems to their specific needs and finding new ways to deliver value.

Additionally, as the "citizen data scientist" emerges, more people actively rely on decades of combined effort and work to create models without ever considering the statistics and infrastructure behind them.

"A citizen data scientist is a person who creates or generates models that use advanced diagnostic analytics or predictive and prescriptive capabilities, but whose primary job function is outside the field of statistics and analytics. Citizen data scientists are "power users" who can perform both simple and moderately sophisticated analytical tasks that would previously have required more expertise."

Carlie Idoine, "Citizen Data Scientists and Why They Matter"

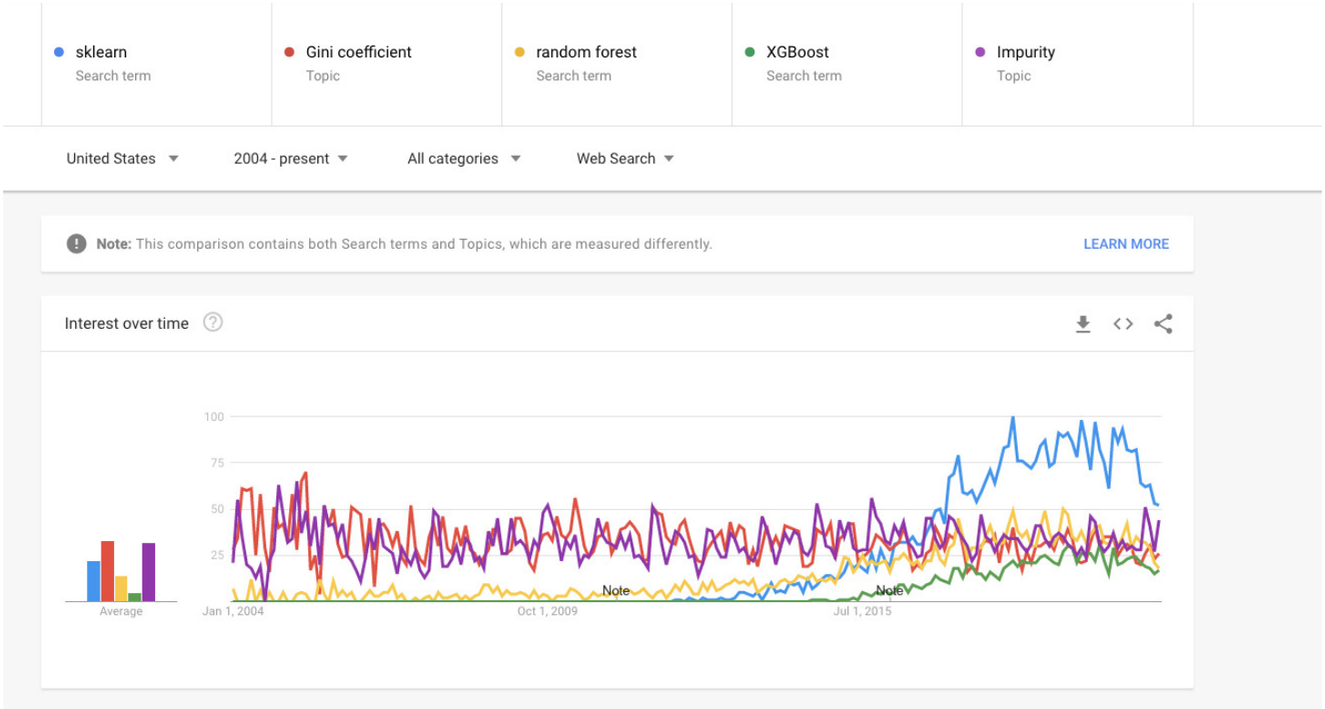


Figure 1: *Trending use of sklearn, an open source library for applied machine learning and data science, suppresses common statistical terms*

The skills you need to succeed have changed as well.

The first wave of data science required deep knowledge of theoretical statistics along with coding experience. That laid the foundation for today's ecosystem. The most essential skill today is resourcefulness. Specifically, the resourcefulness to find and leverage the right ideas and data for organizations' business needs.

With the barriers to access lowered, the greatest obstacle keeping enterprises from unleashing the full potential of AI and data science is **having the right data to feed their models.**

But having the right data isn't a matter of chance. The organizations leading the way have established, consistent, systematic processes to get the data they need.

In other words, they have a defined data acquisition strategy.

Towards building a data acquisition strategy

Forward-thinking organizations understand that having the right data to fit their analytics is vital for success. The next step is to consider the processes you use to acquire data - and to build a consistent, methodical data acquisition strategy.

Why is a data acquisition strategy necessary?

Finding the right data is a complex process. If you need to do it more than once (or continuously) it's simply unsustainable to do this ad hoc.

Maximizing data acquisition efforts takes a coherent approach, right from the point of identifying the data through to controlling costs and checking for quality. More importantly, it requires automation.

Beyond this, there are few other factors that make it vital to have a right data acquisition strategy:

Data is competitive advantage

A bank that can access more data about their business and customers can make better default predictions, target fraud, and generally improve their services to offer a better experience. A retailer with the right data can predict when demand will spike and adjust accordingly. They can also create better-targeted marketing campaigns and offer the right promotions at just the right time.

This is true even if they keep using the same models and analytics tools. **All else being equal, better data leads to competitive advantages.** Better data is the key to winning market share.

Take hedge funds, which understand that the right data generates clearer signals and a significant advantage over rival traders. The industry uses alternative data to better understand historic trends, compare it to present movements, and find potential long-term investments that can deliver big returns. This process, which investors call

"finding alpha", is a staple of the investment world.

But "alpha" isn't just something investors need to look for. Any industry can define their own. It simply means the unique data asset that no one else has - and that creates massive differentiation.

Imagine a CPG organization. It's common knowledge that holidays and weather patterns impact consumer demand, but it's hard to distill or quantify that impact. Much less to use it in predictive algorithms.

A retailer or manufacturer that can properly contextualize this data along with their own would instantly have an advantage over competitors in the form of better predictive capability, and smarter inventory and production planning. To gain competitive advantage, organizations increasingly look beyond their own data assets to access external or alternative data.

"Organizations that stay abreast of the expanding external-data ecosystem and successfully integrate a broad spectrum of external data into their operations can outperform other companies by unlocking improvements in growth, productivity, and risk management."

Mohammed Aaser and Doug McElhaney, "Harnessing the power of external data" (McKinsey)

External data sources are expanding in volume and variety

Data collection rates have increased internally for most organizations, but they're nowhere close to keeping pace with the number of external data sources that become available every day. This goes well beyond volume: there's more variety, too. Of data types, data providers and possible dimensions.

The abundance of data is obviously a great thing - after all, it means organizations can find exactly what they need. But it also complicates matters.

Many organizations struggle to keep up with the pace of growth while seeking out new opportunities and ways to leverage new data. Take the recent explosion of open data sources, for example. Anyone can scour through to find exactly what they need - but the result is a diverse, vast, and often overwhelming data ecosystem.

Meanwhile, more and more enterprises and other organizations are looking to monetize their own data. This includes datasets that may not have internal value, but could be significant for other companies. Again, this opens up exciting opportunities, both for these enterprises and potential buyers of the data. But connecting to and navigating the data ecosystem is an enormous challenge.

For the average organization, keeping up with the pace of growth in alternative data - while finding opportunities and ideas for new, relevant types of data to use in business processes - is far easier said than done.

A Sample of Explorium's Data World
(which will be outdated by the time this article is published)


Foot traffic, or anonymous mobile GPS locations


Professional profiles


Consumer survey


Credit card aggregations


Company profiles


Real estate


Pricing data


Firmographics


Technographics


Marketing attributes


Financial information


Market Value


Web analytics


Employment


Online presence


Reviews

Fear of missing out

What happens if you don't seize the opportunities presented by these ever-growing data sources?

Well, it could mean missing out on a big deal or launching a new product line. It could mean watching your competitors seize market share. Or wasting marketing budget on the wrong leads. Failing to anticipate spikes in demand that cost you dearly.

Enterprises that strive to be truly data-driven understand that ignoring the right data could equal millions in potential revenue.

That's why they develop tools, processes, and methodologies to search, discover, connect to, and acquire relevant external data. Eventually, those building blocks are assembled into a coherent, systematic data acquisition strategy — a process that identifies relevant data, quickly onboards it, and manages data across projects.

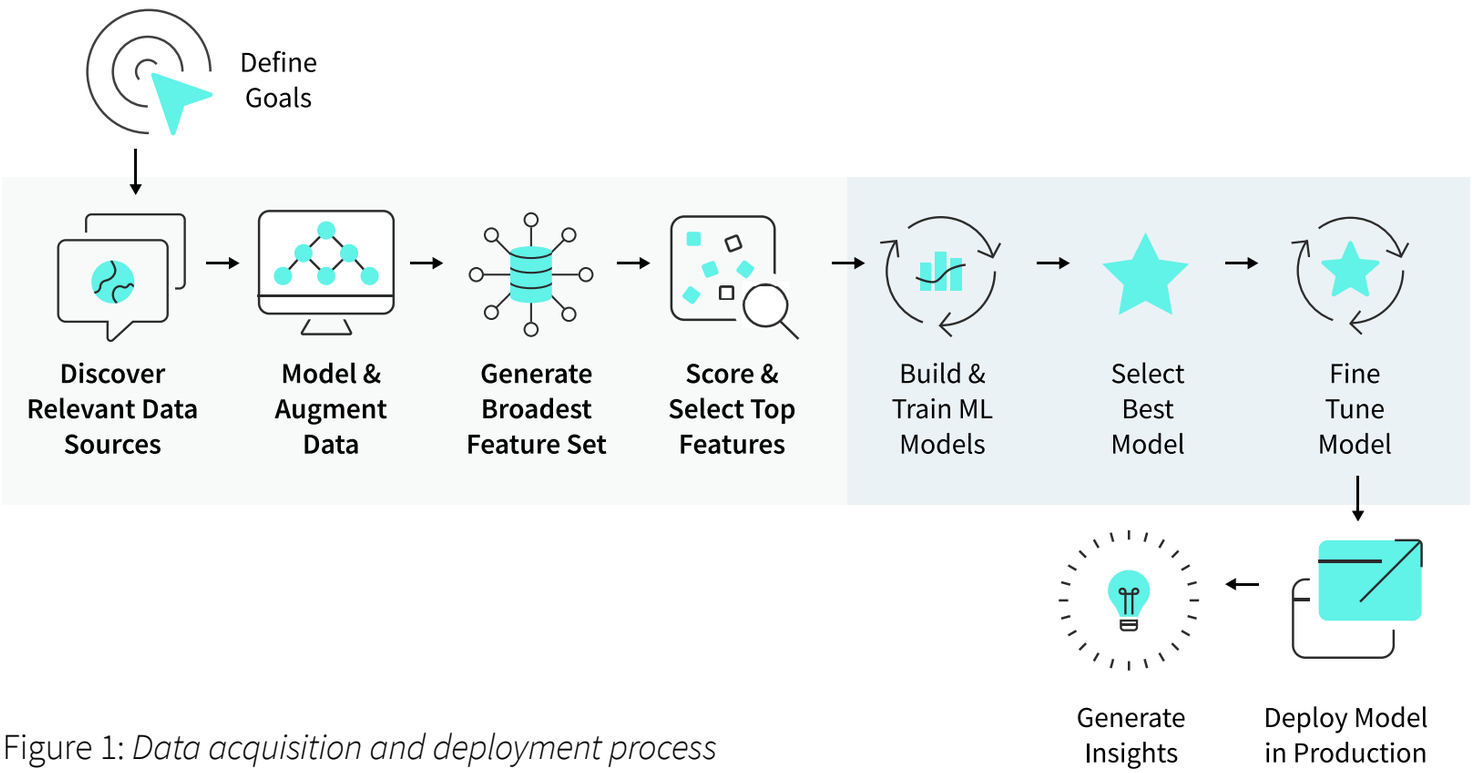


Figure 1: *Data acquisition and deployment process*

Unlocking the power of alternative data

The key components of a successful data acquisition strategy

A data acquisition strategy doesn't happen overnight. Building a "data hunting" organization — one that actively seeks out data opportunities — requires that you put the right resources in place and choose the right teams and platforms to manage them.

Let's take a closer look at how this works.

Streamlined data acquisition

Data hunting only works when the right infrastructure is in place. To truly leverage external data for innovation, experimentation and overall expansion of data assets, forward-thinking organizations must build the following:

1. Standardized search processes
The first step to optimizing data acquisition is to create a specific "data hunter" position, such as a data acquisition manager. This person will be responsible for constantly seeking out new data assets to support new and ongoing projects. Additionally, enterprises should also establish how they qualify and test datasets for coverage, quality, and relevance.

2. Effective procurement processes
Data seeking must also include a procurement process that can keep up with the volume of data you need. Unlike software, which requires a single license and is usually standardized, data often comes with unique terms of use. It's crucial to get these right.

Usually, data procurement occurs in two steps:

- **Pre-decision testing:** The first part is figuring out if the data is actually relevant. This process gets more complicated if you need to negotiate and sign NDAs, DPAs, and other security and privacy agreements. You also need to figure out if it's worth the price tag, which takes a keen understanding of the organization's data needs.
- **Purchasing (Licensing):** Once you've decided that a particular dataset is useful, you can start the actual acquisition. This phase includes agreeing to favorable commercial terms - establishing SLA terms that ensure seamless data integration, legal terms that cover data usage, and compliance requirements.

3. Robust testing protocols

While some data vendors and data sources have truly unique data assets, most data is pretty similar, even if it's sold at a higher price point. For example, we found that in B2B intent data, sometimes a vendor will offer a dataset at 10% of the industry leader's pricing which provides more than ten times better coverage and quality.

This makes it critical that you have an objective, robust testing process for any new data source you acquire and integrate.

Once you've located a promising dataset, test it based on these three factors:

- **Coverage:** The initial testing phase examines whether the data or API you're exploring matches your own data correctly. In demographic data, you might look at whether the source covers the geographic area you're analyzing. For B2B data, you might need the data source to match the companies you're studying.
- **Quality:** Next, you need to understand any gaps and errors in the dataset. How many missing values are there? Are the values accurate? If you purchase a B2B data license only to find that Apple Inc. is listed as having ten employees, that's a pretty big giveaway that this dataset is not up to scratch.
- **Relevance:** Even data with great coverage and quality might not be relevant to you. This part is more art than science, but it's crucial in order to avoid acquiring data you don't need. Let's say you're looking for data to help improve your

underwriting and risk processes. You find a data source that's completely accurate with high coverage, including data on a huge range of companies you're interested in looking at. The trouble is, there's only one data point on each of them: the logo color. Hardly relevant to your objectives, right?

Maximizing data opportunities with automated search and discovery

One of the biggest problems with scaling data acquisition is that too much variety creates paralysis.

Manually conceptualizing, testing, and performing analysis for each project takes too much time and leads to tunnel vision. Overwhelmed analysts and data scientists may start to overlook existing data sources and data opportunities alike. Testing every data asset for every project is simply unfeasible.

… Which brings us to automation.

Organizations facing these challenges must adapt by using technologies that help them optimize the steps. It's essential to find a platform that allows you to quickly discover and test a wide variety of data sources based on each use case you have.

Removing Data Roadblocks: Centralized acquisition

Another major roadblock to data acquisition is that you have to deal with multiple vendors for unique data assets.

Dealing with multiple data sources brings up a variety of issues, both logistical and technical. On one hand, you need to negotiate, agree on, and sign multiple disparate

usage agreements. On the other, you constantly need to worry about integrating every unique data asset into your internal datasets.

A centralized data platform that automates data discovery, matching, and integration can solve this issue. What you really need is a platform that can connect to multiple data sources without needing to integrate with each specific data vendor separately.

This isn't the only reason to use it, though. Opting for a single, centralized, data partner over multiple niche sources delivers a number of other benefits:

Security and compliance

These are two major considerations when dealing with alternative and external data sources.

Most APIs available through premium data sources require inputs that could contain sensitive information (names, emails, locations). This means that, when you're enriching a dataset about individuals using a premium source, you would have to share identifying data to match against the external data.

While there are protections in place in most agreements, having to manually navigate this landscape with multiple providers is a major bottleneck.

Keeping up with the growth rate of data assets

Expanding your data assets rapidly for scaling and variety can cause problems too.

First, the volume of new contracts with vendors is growing, leading to more time spent on legal and procurement processes instead of on using the actual data. Some negotiation processes can take months for a single provider— doing this at scale with tens of vendors would take years for even the most well-staffed legal teams.

Second, every vendor has its own schema, APIs, delivery, and authentication systems. You'd need a dedicated team just to handle integration, version management, and

data monitoring! More importantly, this slows down your research and development teams.

Choosing a central external data platform solves both problems. You're only dealing with a single provider, so you exponentially reduce the negotiation and legal process. It also makes initial onboarding and integration much easier. A single integration eliminates the time spent on connecting data, cleaning it, and standardizing delivery methods.

Foot Traffic, Anonymous Mobile, GPS Locations, Sales, Professional Profiles, Consumer Survey, Credit Card Aggregations, Reviews, Company Profiles, Real Estate, Pricing

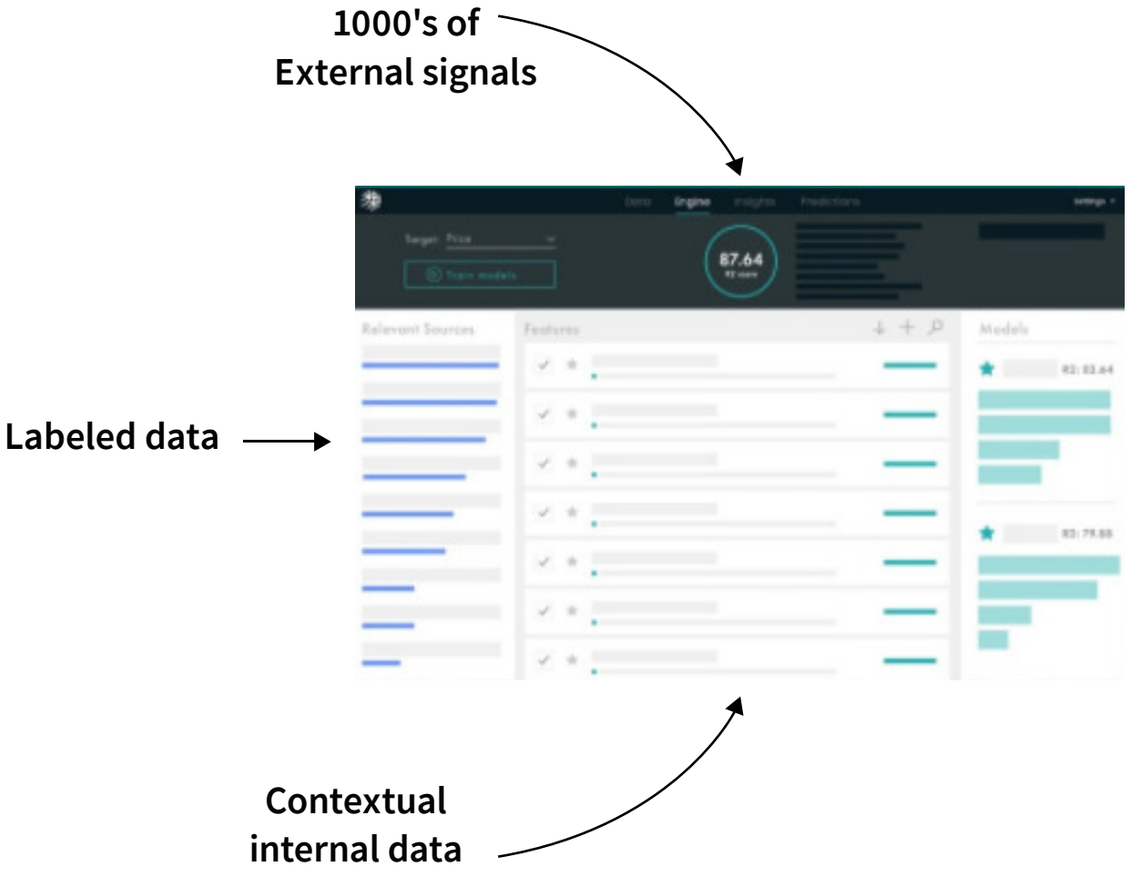
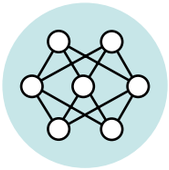


Figure 3: *One place to access all external data*



Other considerations to improve data acquisition

Here are a few other aspects of data acquisition to keep in mind:

Challenges with specific data types

Whenever you deal with unique datasets and data types, you'll inevitably run into different challenges with each of them.

Consumer data, for instance, requires much higher regulatory compliance, due diligence, and dealing with personal identifiable information. Geospatial and demographic data - which is crucial for retailers and CPGs - is almost always open source, making quality the primary concern. The challenge with open source data is that pinpointing the right data amongst everything that's available, while keeping it updated, is like finding a needle in the proverbial haystack.

Choosing the right problems to tackle

Fine-tuning your data acquisition also requires you to understand what types of problems you're trying to solve. Generally, you can break them into two groups — solving a previously "unsolvable" problem, or optimizing your existing core business models.

- **Solving an unsolvable problem:** one benefit of alternative and new data in your models is that it lets you solve problems that were previously intractable. With enough relevant data, solutions are easier to find.

  For instance, a global organization that sells to SMBs might benefit from predicting which potential customers are likely to go out of business in the next year. The quick solution would be to build a predictive model, but which data would you use? First-party data isn't really valuable in this case, as it doesn't account for

or measure external factors. Alternative data, like business financial filings and reports, could provide a much clearer picture.

- **Optimizing existing core business models:** soaring adoption of AI, ML and data science has created a wealth of models companies can use. But these vary in effectiveness - and may are far from perfect.

  Even small improvements can have a huge impact, though. For most organizations, a 1% improvement in performance could have a significant impact on the bottom line. A better credit risk prediction model at a bank that boosts its accuracy by 1% to 2% could result in millions of dollars in savings, for example.

  Usually, achieving these marginal improvements involves heavy investment of resources and time spent fine-tuning models. This results in more complex models that, while definitely better, are harder to read and explain. This makes them less useful in heavily regulated fields, where you need to show how exactly how you reached any given decision.

  On the other hand, additional data sources and signals present an easy win — they let you add new dimensions and extract better insights from the same model. Banks, for instance, increasingly use alternative data for underwriting and risk management. Retailers use data on competition, events and the weather to manage pricing, plan inventory, and optimize their supply chains. ECommerce sites get better personalization and LTV calculations from their data usage. The key factor is that alternative data makes models more accurate, offering a much more scalable improvement.

Addressing the challenges

The question, then, is how to overcome these challenges to make the most of the data acquisition opportunity.

For data teams, this means considering the following factors to find answers to some tough questions:

- **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 by having a team that understands your organization's data needs. One that proactively seeks out opportunities to get better data and better results.

- **Process:** It's not enough to know that you need data and then go looking for it. You need to build a data pipeline. This 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.

- **Technology:** Finally, you need to give your team the right tools and platforms to succeed. Sure, they could deal with datasets and providers as they come, but this will only create bottlenecks. They need to have tools that let them find the data quickly, 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.

Patterns for Successful data acquisition strategies

Data acquisition is a multifaceted and complex process. It impacts teams across everything from data engineering to legal and compliance.

Data-driven enterprises understand this. They anticipate the challenges. They build the right teams and empower them with technology stacks designed to deliver the tools, speed, and processes necessary at every step of the process.

A powerful data strategy — and platform — that optimizes data acquisition contains three key elements:

Access

The first, most obvious way that a data platform provides value is access to data. Not just any data: the right data. Organizations that require alternative data for multiple use cases and business challenges don't have time to constantly hunt through thousands of potential sources just to find the few that are relevant.

A data platform should provide instant access to thousands of potentially relevant data sources and signals, and facilitate the process of finding the optimal ones. Explorium's external data platform offers a data catalog consisting of thousands of premium, proprietary, and open source datasets that are standardized, harmonized, and ready to integrate.

Take Signal Studio, a product designed for data science and business analyst teams. This gives enterprises an easy way to search through the most relevant sources to find

the signals and datasets they need - and integrates them into their existing ML and analytics pipelines. Once your data has been enriched, you can use your domain knowledge and experience to add specific signals, helping you get the most out of your models.

Automated data discovery

On its own, access isn't enough. A data platform needs to make data discovery and matching seamless. The ability to automate the discovery process can save you weeks or even months, expediting your time-to-insights.

Explorium's automated data discovery engine quickly connects your internal datasets to the most relevant data sources. The ones that will give your models the greatest uplift. More importantly, this also automates the continuous data enrichment process, letting you focus on the bigger picture.

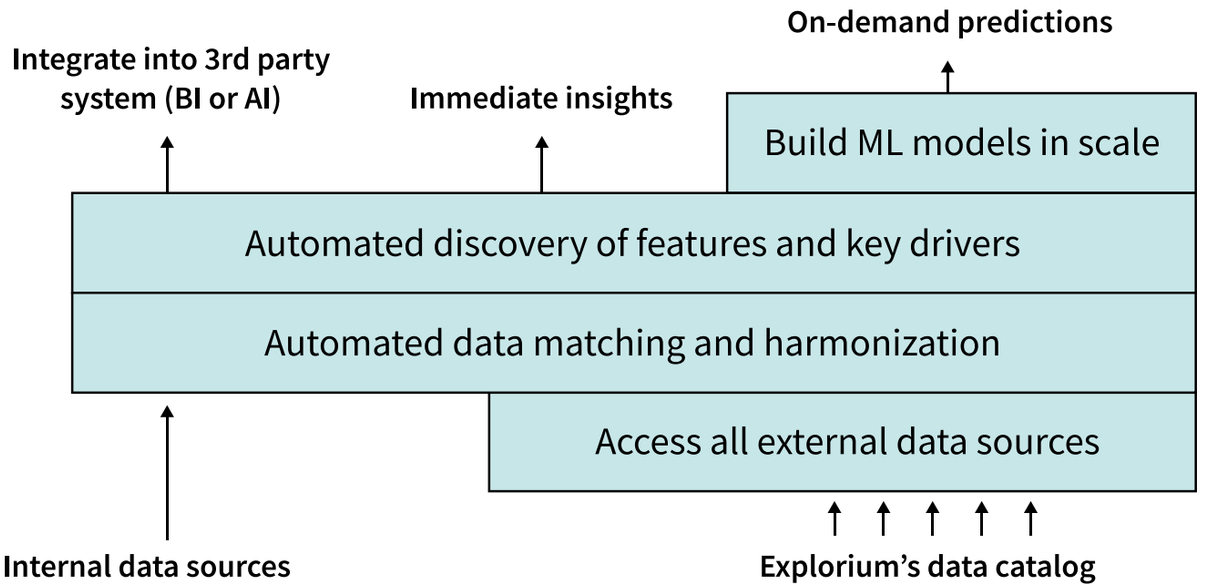
Production pipelines

Even after you've found and integrated your data, you need to keep monitoring it for quality, freshness and relevance. Data or model drift, outdated data, and schema changes can lead to serious issues and impact the effectiveness of a data acquisition strategy. The ability to keep track of issues, data relevance, and any other problems is paramount to a data platform's value.

Explorium automatically ranks each data source, and produces pipelines in runtime to keep up with data needs across use cases. What's more, the platform lets you re-run any dataset to find new, relevant enrichments and features. You can see how datasets impact your models' accuracy, and set up regular updates in advance.

How Explorium Helps

Finding and using the correct data for positive business outcomes is resource-intensive and expensive. Explorium not only helps you identify the right data signals, but also extracts meaningful insights - and makes the whole process easy. You get thousands of data signals brought together, available at your fingertips. You get powerful data discovery tools and feature engineering in one place. All of which means significant, demonstrable ROI.



A single place to access thousands of signals

Explorium's first-of-its-kind, automated, external data platform will connect you to the data sources you need in minutes, not months. It also helps you distill the right features and insights to drive better business decision-making.

Powered by automated data discovery and feature engineering, Explorium immediately connects you to thousands of premium, proprietary, and public data signals. More importantly, it gives you powerful tools to find the answers to your predictive questions. Explorium validates the quality and compliance - so you don't have to. The data you access is always current and compliant.

Faster time to value

Explorium is your one-stop-shop for all external data needs. You significantly reduce the time and money spent looking for the right external sources. You're no longer stuck negotiating with multiple data providers or risking poor ROI of data purchases.

What's more, you get to focus all your expensive data science resources on boosting your predictive models. On incorporating relevant data sources like company, individual, geospatial and time-series data, as well as Explorium's proprietary signals. Explorium automatically matches and integrates external data with your internal data, reducing the time to value even further.

End-to-end platform with augmented feature engineering

Explorium's end-to-end platform harnesses AI-powered data enrichment and feature engineering to help you turn data into powerful insights and business impact.

Your data teams get the best answers to your predictive questions and steer your business to higher revenues. Explorium's powerful feature extraction algorithms let you uncover the right signals to enhance your ROI.

Easy to use

Your machine learning models are only valuable if they deliver on their promise in production.

Explorium lets you build, train, test, and deploy models with just a few clicks. You can now make your data science count by deploying models when they're relevant; when they make the most significant impact to your organization. Explorium's Signal Studio offers citizen data scientists and business analysts an easy way to find the external data signals they need for their initiatives, too.

With Explorium, you get an end-to-end, automated, external data platform for advanced analytics and machine learning that automatically connects and matches your internal data with thousands of relevant external data sets. Explorium accelerates the value from advanced analytics and machine learning models, improving outcomes for complex business problems.

About Explorium

Explorium offers a first of its kind, end-to-end, automated, external data platform for advanced analytics and machine learning. Our unique, all-in-one platform automatically connects and matches internal enterprise data with thousands of relevant external datasets to accelerate the ROI of your ML investment, helping to solve complex problems. The Explorium platform empowers data scientists and business leaders to drive decision-making by eliminating the barrier to acquire and integrate the right external data. It dramatically decreases the time it takes to deliver superior predictive power.

Learn more at www.explorium.ai