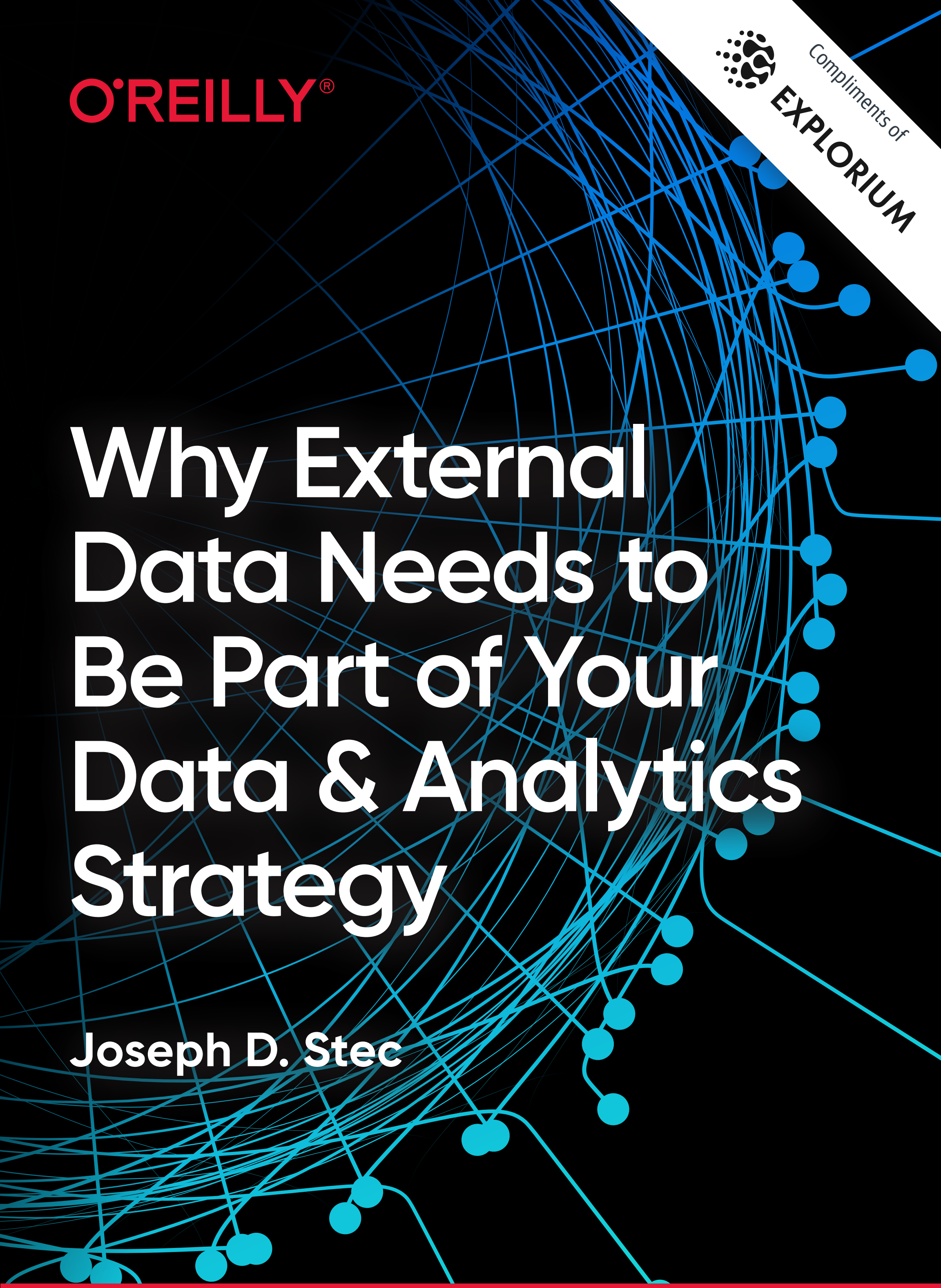
O'REILLY®

Compliments of
EXPLORIUM



Why External Data Needs to Be Part of Your Data & Analytics Strategy

Joseph D. Stec

REPORT

Why External Data Needs to Be Part of Your Data and Analytics Strategy

Joseph D. Stec

Beijing · Boston · Farnham · Sebastopol · Tokyo O'REILLY®

Why External Data Needs to Be Part of Your Data and Analytics Strategy
by Joseph D. Stec

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisition Editor: Michelle Smith
Development Editor: Angela Rufino
Production Editor: Beth Kelly
Copyeditor: nSight, Inc.
Proofreader: O'Reilly Production Services
Interior Designer: David Futato
Cover Designer: Karen Montgomery

January 2022: First Edition

Revision History for the First Edition
2022-01-10: First Release

The views expressed in this work are those of the author and do not represent the publisher's views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and Explorium. See our statement of editorial independence.

978-1-098-12010-8

[LSI]

Table of Contents

CHAPTER 1

Why Use External Data in Your Analytics?

According to Analytics Steps, most companies nowadays are taking advantage of data to improve their competitive position and market response rate. Newsfeeds are awash with stories detailing how retailers, banks, and social media platforms are leveraging data. You can't seem to buy a coffee, post on social media, or listen to your favorite song without a company asking for your personal details. Global giant Spotify, for instance, delivers music to listeners around the world and openly uses internal data for the following purposes:

- Marketing, promotion, and advertising campaigns
- Feature development and evaluation
- Business planning, reporting, and forecasting
- Fraud detection and prevention

However, Spotify also uses third-party or external data to deliver more relevant advertising to its listeners. Spotify's ad partners help them facilitate tailored ads to match your interests or moods, such as which cars a car lover might want to know more about.

Today, innovative organizations—those already using advanced analytics software powered by big data—are like Spotify. According to TechTarget, they're combining data from a variety of internal and external sources to enhance customer service, boost sales, make marketing more efficient, enhance products and services, and infuse more real intelligence into their operations.

Instead of simply deriving static reports from data that has been moved in and out of data warehouses, clever companies are using advanced analytics tools that can simultaneously collect, mix, and match diverse data from disparate data sources in order to improve products and brand loyalty, generate better conversions, identify trends earlier, and pinpoint additional ways to improve overall customer satisfaction.

According to Jennifer Belissent's Forrester blog, the organizations that can create better infrastructure to collect, store, analyze, and leverage external data—and successfully integrate it into their operations with their internal data—can outperform other companies by unlocking improvements in growth, productivity, and risk management.

This report expands on the aforementioned points and answers the following questions expounding on the rise of external data:

- How is new technology making external data easier to use with analytics?
- How does an external data platform fit into your data architecture?
- How can you start leveraging external data today?

Fusing Internal Data with the Right External Data

Footfall traffic is a good example. Footfall is how retailers describe the number of customers who enter their stores. Cuebiq explains footfall attribution, a related concept, as a method used to correlate digital marketing campaign impressions and actual store visits.

According to Knorex, footfall attribution is essentially an ingenious mix of mobile campaign impression results and the data collected from actual store visits. Instead of just relying on mobile marketing techniques, cafes, restaurants, supermarkets, and various retailers can use footfall attribution to gain valuable insight into competitive analysis, temporal analysis, and customer analysis. Most importantly, they can measure what matters—their precise sales growth.

While footfall traffic is an extremely effective approach for helping retailers and restaurants monitor and grow sales, it is only the beginning of the possibilities with external data. For example, the *Wall Street Journal* reported that retail giant Canadian Tire delivered a 19.1% increase in year-over-year retail sales through Q3 2020 by mixing and matching foot traffic, weather, traffic patterns, and shifts in demand for bicycles and outdoor furniture.

Hershey, meanwhile, was better able to pivot its supply chain during the 2020 lockdowns by understanding the different types of chocolate that were being consumed at home. According to the *Wall Street Journal*, this directly led to a 5.5% increase in same-store sales.

Tom Davenport, author of *The AI Advantage*, underscores the ever-expanding importance of fusing internal data with the right external data, when he wrote for *MIT Sloan Management Review*:

> Trying to model low-probability, highly disruptive events will require an increase in the amount of external data used to better account for how the world is changing. The right external data could provide an earlier warning signal than what can be provided by internal data.

Data Hunters Know That Quality Always Outfoxes Quantity

As decision makers are coming around to understanding the value that seeing the big data picture brings to their respective organizations, they are working harder than ever to find the correct data sources that will give them an edge. Deloitte Insights notes that 92% of data analytics professionals said their firms needed to increase the use of external data sources, while 54% said their company plans to increase spending on it. As noted in Belissent's blog, the former chief data officer (CDO) of Flagstar Bank acknowledged, "With our own data, we can only look internally. We need to see industry benchmarks, regional trends, what waves we can ride on; we derive competitive advantage by getting data from outside and enhancing our own data."

This new quest for the Holy Grail of data—to get a leg up with an inside scoop or unique source of information or local knowledge—has led Forrester to the following conclusions:

The demand for external data is increasing in parallel with firms' abilities to source it.

Companies are already well aware of the need to better leverage internal data such as transaction data, customer interactions, and other process and performance metrics. Yet, there is more need now to supplement their internal data with external data such as weather, traffic, social media listening, partner data, and economic data from third-party sources.

The supply of data is keeping pace with demand.

As more companies are looking for external data, data supply and data sources have accelerated. Traditional data providers are issuing new data offerings as "originators scrape websites for pricing and product information."

The number of firms commercializing data is mushrooming.

Data marketplaces have been popping up all over the place. Data commercialization is also on the rise, with companies developing new data-fueled products and services and data brokers assisting clients to spot new sources and even host events for niche data to meet potential buyers.

The rise in demand for external data has fostered a new employment role and changed the face of mergers and acquisitions. According to InformationWeek, data hunters—also known as data acquisition specialists and data scouts—are now in high demand, and, according to Nextgov, companies are creating new positions solely to find the best external data. As Econsultancy reported, Swiss pharma giant Roche's acquisition of the health technology company Flatiron Health, which helps collect data that could be used for cancer research, also shows how far companies are willing to go to get the best third-party data.

How Does External Data Improve Analytics?

In this section, we explore how external data elevates and enhances how organizations can analyze and interpret data outside of their apps or databases.

Learning About the Power of External Data from Your Next Dinner Delivery

If you're working from home and starving for lunch, you're more likely to stay inside and order takeout if it's raining outside. Any one of the thousands of food delivery drivers across America knows that bad weather leads to delivery spikes. But now, drivers for Deliveroo have at their fingertips the external data to back up this claim, and they can also instantaneously share changes in the weather or traffic with fellow drivers.

For online delivery platforms such as Deliveroo, leveraging external data is becoming paramount in an industry with stiff competition and razor-thin profit margins. Meanwhile, DoorDash, a leading online delivery platform, is looking to upgrade its data architecture to access geospatial data to better understand the economic impact of a store's location, according to Snowflake, "to analyze different configurations and extend the geometry to influence supply and demand."

Another online food ordering company, Grubhub, studies external data to understand the loyalty of its customers. Grubhub believes its online diners are becoming "more promiscuous," that is, the company is concerned that its newer diners are increasingly coming to Grubhub after already having made orders on a competing online platform.

Grubhub's external data indicates that its existing diners are increasingly ordering from multiple platforms and find that this so-called "platform-sharing" is most common among its newest diners and markets. The trend, however, is also spreading to Grubhub's core diner base.

For instance, according to Bloomberg Second Measure, 61% of Grubhub customers did not use another meal delivery service in the second quarter of 2019, but that number fell to 46% two years later. DoorDash, conversely, had 58% of its customers using them exclusively in the second quarter of 2021. By harnessing external data, Grubhub now realizes that the "easy wins in the market" are quickly evaporating.

Data Is Food for AI, So Don't Feed It Junk

Computer scientist and technology entrepreneur Andrew Ng recently went on record to emphasize a shift toward a data-centric approach to machine learning and AI. He explains the importance of companies using the correct data over simply using *more* data. His message is straightforward and might even be kept in mind the next time you decide to use a meal delivery service: data is food for AI, so don't feed it junk. Ng makes several sobering observations:

- He notes that 80% of the effort on a machine learning/AI project is spent on preparing the data, and only 20% on modeling. Despite this fact, 99% of AI research focuses on model-centric approaches to improving results.
- The most vital task of MLOps is to ensure consistently high-quality data in all phases of the machine learning project lifecycle.
- Cleaning up labels (making them more consistent) is a more efficient way of improving accuracy than collecting more data, especially for small datasets (<10,000 observations).

Ng reiterates that big data should focus on improving data rather than model accuracy: "Now that the models have advanced to a certain point, we have got to make the data work as well."

In an *ACM Transactions on Computer-Human Interaction* article, Google researchers support Ng's claim:

> Paradoxically, for AI researchers and developers, data is often the least incentivized aspect, viewed as 'operational' relative to the lionized work of building novel models and algorithms. Intuitively, AI developers understand that data quality matters, often spending inordinate amounts of time on data tasks. In practice, most organizations fail to create or meet any data quality standards, from under-valuing data work vis-a-vis model development.

Nothing threw big data's model-centric approach into question more than the COVID-19 pandemic. A recent McKinsey article states:

> In a few short months, consumer purchasing habits, activities, and digital behavior changed dramatically, making pre-existing consumer research, forecasts, and predictive models obsolete. Moreover, as organizations scrambled to understand these changing patterns, they discovered diminishing value in their internal data.

> Meanwhile, a wealth of external data could—and still can—help organizations plan and respond at a granular level.

According to Forbes, advanced external data may include brand loyalty on social media, real-time product information (price, discount, stock status, etc.) in ecommerce marketplaces or competitors' websites, and suppliers' information tracking. Kabbage is an example of a fintech company taking advantage of external data. Kabbage determines eligibility for issuing loans—and the terms under which a business would pay it back—by tapping a vast variety of sources, from traditional accounting statements to social media signals. The social media company then loads this data into its proprietary machine learning algorithms.

The report from Deloitte Insights notes multiple other examples of analytics programs generating value from external data, such as helping businesses personalize marketing offers, enhancing HR decisions, acquiring new revenue streams by launching new products or services, enhancing risk visibility and mitigation, and anticipating shifts in demand more precisely for their products and services.

Angie King, principal at End-to-End Analytics, recently told the *MIT Sloan Management Review*: "The benefit of using external data is so great that there are businesses built around gathering this data, consolidating it, cleaning it, and packaging it up for use by other companies." King also noted that using external data can improve a company's predictive analytics and machine learning models: "Without having external data capturing these events, the predictive models wouldn't be able to infer the reason for the resulting spikes or drops in sales."

Whether it's an agro corporation using geolocation and weather data to help a farmer, a bank accessing social media to determine credit worthiness, or a logistics manager using a news feed to determine potential supply chain disruptions, companies should start identifying the best new technology that is making external data easier to use to augment their analytics and machine learning models.

Differentiating Data Providers from Data Marketplaces

While collecting the correct data is paramount, timely, relevant, and high-quality data remains elusive. A Forrester Consulting study found that 99% of firms surveyed faced issues with customer data, while 96% indicated that timelines and accuracy issues with acquisition of customer data were big problems. Determining what data is needed before purchasing it can be difficult, and pinpointing up-to-date, high-grade data is equally challenging. This problem has led to the emergence of data marketplaces.

Data marketplaces are online platforms that facilitate the buying and selling of datasets from several different sources. Data marketplaces are usually cloud services where individuals or businesses upload data to the cloud and provide self-service data access while guaranteeing security, consistency, and high quality of data for both parties.

Data marketplaces also facilitate data monetization. An AI software platform that wants to train and sell its AI-based models, for example, could purchase data from a marketplace. Data marketplaces include personal, B2B, and Sensor/IoT (the Internet of Things) data, and offer the following types of data:

- Business intelligence (BI)
- Market research
- Geospatial
- Demographic
- Firmographic
- Public

While data marketplaces seek to build an ecosystem of data providers and data consumers by providing data access, purchasing a dataset is no guarantee of a specific business outcome. Buying datasets can also come at a very steep price, as licenses can cost as much as hundreds of thousands of dollars.

Here are some other limitations of data marketplaces:

- Integrating external data can be a costly challenge requiring separate tools, platforms, or data science teams.
- Purchasing, formatting, handling, managing, and integrating external data doesn't guarantee ROI.
- Locating the most robust, applicable data isn't always clear.
- Matching the format that the organization's data is in can be a serious setback.

- Adhering to security and compliance regulations such as General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) can become a major headache.

Understanding which data is best for your business needs and then unlocking and integrating it with your own internal data is where an external data platform steps in. An external data platform can help with every step of the data acquisition process—from data discovery to data prep, integration, model training, compliance, deployment, and model retraining. An end-to-end external data platform can provide access to all of the relevant external data sources in one platform, allowing you to understand which data signals you need and how they will drive ROI.

What Are Data Signals?

Data signals are pieces of data that help you understand and contextualize the experiences or situations of your audience, clients, partners, or patients. Common data signal categories include:

Company data
: Basic data like industry and North American Industry Classification System (NAICS)/Standard Industrial Classification (SIC) codes, economic indicators like revenue trends, technologies, reviews, and more

People data
: Contact information, social networks, interests and hobbies, purchase habits, and more

Geospatial data
: Demographic data, footfall traffic and trends, and other indicators

Temporal data
: National and local events, weather indicators and history, etc.

Product data
: Pricing data, ratings, and retailers, etc.

KPMG notes how the three following data signals have specifically been used to help healthcare during the COVID-19 crisis:

- Static (e.g., the geographical distance between a home and a hospital)
- Slow-moving (e.g., the ratio of health-care professionals to people in an area)
- Fast-moving (e.g., first-time unemployment filings)

These signals showed which locations had a predisposition for a larger and longer sustained impact from the pandemic and—by extension—more lasting macroeconomic effects. An example of a macroeconomic effect would be if people in a specific area rely on mass transit or if the density of people in the same area has past diagnoses of certain relevant diseases. KPMG is also adding a vital fourth new category: pop-up data. The most recent number of confirmed COVID-19 cases reported in a specific city or county serves as a relevant example.

KPMG also notes that—by leveraging all four varieties of signals—insurers and lenders can now look into and explore and mitigate risk at an extremely local level to help business leaders predict which potential customers carry the greatest risk of mortgage default or insurance claim submission.

Marketers, meanwhile, can implement external audience data known as experience signals that come from different systems, channels, and in-house technology. Signal types include digital clickstream data, ecommerce information, POS (point-of-sale) data, call center interactions, CRM data, service interactions, IoT data, HR data, sentiment captured from videos, sales and marketing tools, and even survey data.

"Spray and Pray" Marketing Methods Are Being Replaced with More Personalized Interactions That Use Relevant External Data

Marketing methods such as surveys, however, may be going the way of the horse and buggy. A recent Deloitte report reiterates that even though customers are increasingly frustrated by such generic offers, most marketers continue their "spray and pray" mass marketing techniques and show little sign of changing. Today's customers, however, only want interactions that are relevant, personalized, and based on a consumer's situation and preferences.

The casino industry, for example, is now revisiting how third-party data can supplement traditional first-party data and gaming metrics. Casinos are now looking at three key factors:

- The growing importance of nongaming spend
- The rapid growth of digital (accelerated by the pandemic)
- The rapidly rising consumer expectations around marketing personalization

Casino marketers can now dive deeper into what their guests and prospects are consuming online and which intent signals they're exhibiting, such as other sites visited, common online transactions, interests and hobbies, and countless other individual variables. Casino marketeers can now create an experience and message that matches real consumer needs.

Casinos also now understand that they're competing with travel brands across a consumer's share of travel. With the integration of third-party data, casino marketers are able to see what other brands their customers are engaging with and use that knowledge to help them understand their broader competitive set.

What Are Data Features?

Data features are specific variables that make up a dataset. The most common features, or measurable pieces of data, include name, age, nationality, race, height, weight, and sex. A feature's data type can be determined to be a percentage, a category, a number, a date, etc.

Determining the correct feature depends on which business problem you would like to solve and what your business goals are. Even within the same industry, different businesses require different features. However, displaying too much information can divert focus away from the essential metrics, and overloading an analytics model with unnecessary features can decrease the accuracy and negatively affect the model's efficiency. This is where feature engineering comes into play and ensures that attributes relevant to the business problems are the only ones selected and fed into the analytics model.

Choosing the correct features will greatly enhance the efficacy of your machine learning model. At the same time, intelligent feature engineering optimizes models by selecting only the relevant variables, thereby reducing the effort to retrain a model if new features are added later on.

CHAPTER 2

How New Technology Is Making External Data Easier to Use with Analytics

In this section, we begin to explore the nuts and bolts of external data platforms solving key challenges, such as:

- Distilling the unknown unknowns while still understanding compliance risks
- Overcoming data access and data usage issues
- Augmenting data for quicker time to value

How Do External Data Platforms Help Find the Unknown Unknowns?

Sometimes you don't know which data you're looking for when you're looking for data. Counterintuitive data signals are often business questions that you can get your head around but can't answer since you're missing the right data. The missing data might be a spreadsheet on someone's monitor halfway around the world. It might be another external data source that will simplify your analytics and help you make a more intelligent business decision. You may not know where the data is, what type of data it is, or what it'll look like—but you ‘ll intuitively know it when you see it. What you've been looking for is what Donald Rumsfeld coined as *unknown unknowns*.

Learning that a business process error has been keeping customers from completing their orders for over a week is an example of an unknown unknown. In a high-risk, high-reward industry such as online lending, unknown unknowns can put you at a disadvantage—not to mention your competition may be using it to gain an advantage and take market share.

The Federal Reserve report "Interagency Statement on the Use of Alternative Data in Credit Underwriting" highlighted the use of alternative data and drew a clear line in the sand: either start leveraging alternative data or start preparing to be left behind. The report also serves as a reminder that great responsibility should accompany great power.

The report details the following developments:

- Banks and nonbank financial firms are using or contemplating a broad range of alternative data for use in credit underwriting.
- Federal agencies recognize that using alternative data may improve the speed and accuracy of credit decisions and may help firms evaluate the creditworthiness of consumers who currently may not obtain credit in the mainstream credit system.
- The continuing evolution of automated underwriting and credit score modeling offers the potential to lower the cost of credit and increase access to credit.

How does this change the way banks, online lenders, and fintech companies implement alternative data to assist their fraud detection, account management, and credit writing? In the past, a credit underwriter might only take a look at traditional data on a loan application such as income or revenue or past credit repayments. Now, for example, underwriters will look at website registrations, online reviews, and other online alternative signals to determine the creditworthiness of small business loan applicants. The same principles can be applied to consumer marketing.

If you want to augment your approach to catch fraudulent loan applications or loan stacking, machine learning models would be helpful. Your in-house data alone, however, might not optimize the effectiveness of this approach. With the right external data platform, you could connect to the following alternative data sources:

- Government filings
- Business registrations
- Social media posts
- Domain information
- Search engine results
- Foot traffic

Now you can analyze all the data in a single, comprehensible depository and sidestep the laborious data acquisition and data-matching processes. From there, you can pinpoint suspicious patterns, red-flag anomalies, and fraud with machine learning algorithms and turn away fraudsters and unqualified customers.

Maintaining Data Quality and Quelling Compliance Risks

While identifying the requisite unknown unknowns is vital, organizations still must maintain data quality, protect consumers, and meet compliance regulations. Strict rules, such as the California Consumer Privacy Act (CCPA) and the EU's General Data Protection Regulation (GDPR), govern how organizations collect and use people's data. GDPR fines increased by 39% in 2020 compared to the previous 20 months, according to DLA Piper, a global law firm specializing in privacy and data protection.

Proper data privacy compliance involves identifying, classifying, and documenting internal and external personal information. The GDPR requires that businesses are obligated to correct inaccurate or incomplete personal data, yet many organizations neglect the importance of data validation. Data quality measures the completeness, accuracy, and timeliness of enterprise data. Without comprehensive data quality controls, organizations cannot locate and resolve data inaccuracies involving personal data.

Unfortunately, resolving important data quality issues does not guarantee compliance. Instead, organizations must eliminate all siloed data tasks by integrating data quality efforts with data governance and data catalog initiatives. The aforementioned Federal

Reserve report on the use of alternative data in credit underwriting underscores the importance of sound compliance management to keep pace with the new regulations:

> A well-designed compliance management program provides for a thorough analysis of relevant consumer protection laws and regulations to ensure firms understand the opportunities, risks, and compliance requirements before using alternative data. Based on that analysis, data that present greater consumer protection risks warrant more robust compliance management. Robust compliance management includes appropriate testing, monitoring, and controls to ensure consumer protection risks are understood and addressed.

As regulations grow more complex and data environments swell, new technologies help maintain compliance. External data platform Explorium, for example, recognizes the essential importance of security and compliance. It understands that a centralized data governance framework empowers a unified approach and promotes collaboration and shared accountability of enterprise data. Compliance measures regarding customers and their data—from the secure software development lifecycle (SSDLC) of the Explorium External Data Platform and data services to the implementation processes—ensure that Explorium protects customers' privacy.

The platform complies with comprehensive, industry-leading standards, regulations, and frameworks and implements the following measures:

- Rest and transit encryption
- Penetration testing
- Risk assessment and management tools applied across the entire scope of the Information Security Management System (ISMS), Privacy Management Systems (PIMS), and Quality Management Systems (QMS)
- Top-tier cloud provisions, such as AWS, Azure, and GCP
- Vulnerability scans, awareness and training programs, and secure coding

The right technology and approach can enrich the quality of your data pool, bringing extensive benefits to you and your customers while still playing by the rules.

Overcome Data Access and Usage Issues

According to Snowflake, the biggest challenge with data sharing—until recently—has been moving ever-increasing volumes of data. For both internal and third-party sharing, data movement, usually via resource-intensive ETL (Extract, Transform, Load) processes, has inhibited the wide-scale sharing of data.

More efficient data sharing improves business efficiency. By sharing live data with internal and external business partners, organizations can optimize spend, provide superior customer service, and streamline operations.

Lydia Clougherty Jones, senior director analyst at Gartner, believes that organizations continuing to focus on getting the right data—not just the most data—to make the best business decisions will vastly improve the quality of their data with improved data sharing: "Data sharing is the way to optimize higher-relevant data, generating more robust data and analytics to solve business challenges and meet enterprise goals...D&A leaders who promote data sharing have more stakeholder engagement and influence than those who do not." Clougherty Jones also understands, however, that there are a few roadblocks hindering data sharing, including:

- Problems with data management
- Insufficient tools and technologies
- Perceived regulatory prohibitions and security risks

While these barriers exist, Clougherty Jones sees tremendous strategic planning opportunities for external data:

> By 2023, 85 percent of data sharing strategies that include external data sources will drive revenue-generating digital business outcomes rather than data management functions, making data sharing an essential business capability.

If more experts such as Clougherty Jones are underscoring the value of external data, what is restraining organizations from taking advantage of this vast resource? Access to external data tops the list. Deloitte Insights identifies several of the key business challenges to accessing external data:

- Determining if continuous data access is needed to refresh machine learning models
- Dealing with usage restrictions
- Understanding if vendors want a share of the revenue the data generates
- Navigating a large, complex data-provider market
- Negotiating purchase and liability terms
- Managing data-provider relationships

Overcoming access issues can be a resource-heavy process and does not guarantee correct insights. Procurement for one data source can take months. Now think about a project requiring 20 to 30 separate data sources. Choosing the right external data platform, however, unlocks instant access to the premier, proprietary, and public datasets. A proficient external data platform should provide access to the most relevant data sources by connecting you to people and to geospatial, time-based, and external company data. A third-party data management solution should not only provide access to data but automatically show you which data sources and features will provide the best model uplifts.

While accessing data poses problems, using the acquired data can present another obstacle for many organizations. On the technical side, according to ProjectPro, the mundane labor of data preparation still accounts for 50 to 80% of a data scientist's daily work. In addition to data prep, organizations must assess the data quality and accuracy of external data, pinpoint inconsistencies between external and internal data, and securely store and catalog the data in an accessible manner—especially taking into consideration that many information management systems, according to Deloitte, were designed to only handle only internal data.

Consistency is the key to properly using data. External data must be matched with internal data to achieve alignment. If your marketing team is analyzing an industry, you must be sure that companies and zip codes match. If, for example, operations are using time from various sources, the times must have the same precision (e.g., milliseconds, nanoseconds, etc.) and format. A complex format such as a protocol buffer, a free and open source cross-platform data format used to serialize structured data, presents more challenges as each source must be updated with the latest definition of the structures. Companies must also be sure their data is not out of date when managing predictive models.

Organizations may still struggle to integrate data into their production pipelines—not to mention the associated costs and complexity—even after prepping and matching external data with their internal data. Avoiding data drift (or unexpected changes to the input data) requires monitoring and maintenance to ensure a predictive model's accuracy. Seasonality, or the predictable changes that take place every calendar year, is a classic example of data drift. For example, retail sales usually peak during the Christmas season and then drop off after the holidays.

Augmenting Data Delivers Quicker Time to Value

We've discussed finding the unknown unknowns in data, understanding data compliance risks, and overcoming data access and usage issues. Now, let's dive into how external data platforms deliver quicker time to value by augmenting your data.

Many business analysts look for data that might be relevant to your use case. A high-level automated external data platform puts you in the driver's seat by asking which data could be relevant for your use case—thus accelerating time to value by infusing top-notch external data into your analytical models to heighten accuracy. For example, a marketing firm might rely heavily on internal data such as:

- Click-through ratios and ad clicks
- Website views, downloads of white papers or ebooks, and registrations for webinars
- Engagement level on social media or other platforms

This internal data is a great start, but it has vast limitations—and thus an adverse effect on converting leads. The same company could augment its internal data by connecting to an external data platform's Enrichment Catalog. This would allow them to ascertain the following metrics:

- Social media interactions with the product and others in the category
- Number of previous purchases in the same category
- Spending potential and financial stability metrics
- Demographic data including cohort group preferences
- Search engine queries in related fields
- Online purchases of more than $30 in the previous six months

By connecting the original internal data to hundreds of external data inputs, Explorium was able to automate the process of collecting signals and quantifying their respective relevance. These signals then fuel the appropriate models to deliver more accurate and robust business decisions. In this case, it led to the marketing firm fattening its pipeline and improving conversions by 18%. By using Explorium's Enrichment Catalog, the firm targeted better leads, improved conversions, and drove higher revenues. Instead of missing out on unknown unknowns, the bottom of the funnel blossomed into a cornucopia of conversions.

Making external data easier to use with analytics starts with harnessing the appropriate data signals to rein in the optimal unknown unknowns. Organizations must then find the most efficient way to access and use the correct data—all while maintaining data quality and adhering to compliance regulations. Finally, leveraging the right external data platform delivers quicker time to value by augmenting the data that has been carefully chosen.

CHAPTER 3

How an External Data Platform Fits into Your Data Architecture

Many data analysts, data scientists, citizen data scientists, and inhabitants of the data blogosphere are familiar with Doug Laney's "3 V's" of big data:

Volume
Refers to vast amounts of data, which can be generated, for example, from cell phones, social media, and photographs

Velocity
Measures the speed at which this vast amount of data is being generated, collected, and analyzed

Variety
Describes the different types of data—structured data (data that can be properly displayed in a data table such as name, phone number, ID, etc.) is blended with current data, which is mostly unstructured: images, audio, social media updates, etc.

Laney recently discussed the possible addition of two more V's: veracity—or correctness and accuracy—and value. In order to improve accuracy and deliver value, data must be processed in a timely fashion, cleaned and stored for analytical purposes, and be monitored by proper governance all while meeting compliance standards. This is a tall order that also requires the infrastructure of scalable processing systems.

The new demands require new solutions, yet legacy business intelligence and data warehousing system architectures—which take both internal and external data from various structured sources—are falling behind. These systems are limited in their capacity and scalability mainly due to being on premises and having been created only for structured data use cases.

Organizations need architecture that can support modern business use cases and solve specific business problems, and modern data platforms should power value creation in a simplified manner for those consuming and producing data. The right automated external data platform can help organizations enrich models with premium external data that connects seamlessly into a modern data architecture.

In this chapter, you'll learn about the components and characteristics of modern data architecture and the evolution of the modern data stack. Let's explore how an automated external data platform fits into the modern data stack and expedites the process of enriching your internal data with the right external data.

The Evolution of Modern Data Architecture

In 1970, former Royal Air Force pilot Edgar Frank "Ted" Codd published a seminal paper, "A Relational Model of Data for Large Shared Data Banks", which described a relational modeling procedure. The result? Codd's approach successfully created database structures that streamlined the efficiency of computers. No longer was data arranged hierarchically. Now, searches could start in more general categories and then be refined into progressively smaller hunts. The new relational approach allowed users to store data in a more organized, more efficient manner, and the science of structured data architecture was born.

Today, a modern data architecture (MDA) can be described as the process of standardizing how organizations collect, store, transform, distribute, and use data. It should also embody the following characteristics:

- Deliver relevant data to people who need it, when they need it, and help them make sense of it.
- Empower decision makers with a user-driven experience that pools the most relevant data to meet business objectives.
- Eliminate silos by combining internal data from all parts of the organization along with external data as needed. In such an architectural utopia, data is neither bartered among business units nor hoarded, but is seen as a shared, companywide asset.

Building the Modern Data Stack

So how does an external data platform fit into the modern data stack? A data stack is made up of all the systems required to facilitate a smooth data journey from inception to disposal. For example, your marketing stack might include CRM (customer relationship management), marketing automation, and analytics tools. Information runs through a stack according to your internal processes and makes data more digestible; it's been said that a data stack is like a kitchen for data and is similar to baking a cake. Most of the ingredients themselves wouldn't be too tasty on their own—unless you enjoy shooting vanilla extract or downing tablespoons of unsweetened cocoa powder.

But a chef with the proper ingredients and a well-equipped kitchen can use bowls, blenders, pans, spoons, and spatulas and turn butter, sugar, and eggs into the centerpiece of your next party. An excellent data stack can turn scattered bits of data into a similar masterpiece, replete with useful fact and dimension tables with clear field names and types—easily digestible by different departments in a company.

Many analytical chefs also understand that they need to add a dash of external data to their next recipe. They recognize that external data provides important context not always captured in internal data: economic trends, consumer preferences, weather, reviews, social media trends, competitive intelligence, and more. Unfortunately, many kitchens are falling behind and are in desperate need of being remodeled and modernized.

The old acquisition process for external data looks something like this:

- Manual data search
- Validation
- Procurement
- Integration
- Monitoring and maintenance

No data provider can guarantee 100% coverage for every dataset, and you may need five to six different providers for a single use case. You may also find 10 to 15 providers claiming that they have the same data with the same quality. Even if you find the best data provider that can enrich your data, this is only the first step of the process.

Next, you need to validate the data but will only be privy to a small sample. This sample must be used as a proxy to ensure ROI—and that is risky. After that, the data must be purchased. Most data scientists and analysts have neither the time, inclination, nor skill set to procure data. They must go to another department—further delaying the process and bloating the budget. This process can take months. Only then can the complex integration process begin.

If you wanted to get daily footfall traffic, for example, you would get three to four terabytes of data per diem—this requires an entire team of data engineers to transform the data and adds another layer of time and risk. Finally, you need to monitor for data drift and make sure the data is up to date as you integrate it into your analytical pipeline. There is also the possibility that the original data sample was not accurate and you might have to start the whole process over again.

The right automated external data platform offers an alternative to this tedious process and can be a one-stop shop for all of your external data needs. A one-stop data shop, which will be covered in further details in the next chapter, can provide the following features:

- Easy access to external data all in one place
- An organized data catalog that provides instant access to datasets and is accessible to all members of an organization

- Seamless integration into BI and ML processes with several ways to integrate data into your analytics pipeline
- A guarantee of data quality and consistency
- Recommendations for the best external data to add to your ML model to improve accuracy

If any data practitioner—be they data analysts, business analysts, machine learning engineers, data engineers, or data scientists—needs external data, they no longer need to jump through the five aforementioned hoops.

As you can see, data architecture has come a long way in the last half-century. We've gone from groundbreaking, yet rudimentary, relational modeling procedures to the 3 V's. We've also journeyed from only accessing structured data to the tedious process of searching for, sampling, procuring, and integrating external data—all the while with uncertain ROI.

Data architecture has evolved to the point where not only is the kitchen modern but its chef no longer needs to drive all over town to find every vital ingredient. Now, the right automated external data platform allows any chef to easily go online and easily access all of the requisite data ingredients at the click of a button.

CHAPTER 4

Getting Started with External Data Today

As the COVID-19 pandemic grew worse, the rapid changes in the global economy and seismic shifts in consumer behaviors rendered thousands of datasets and analytical models useless overnight. Organizations worked tirelessly to find and source more relevant public and third-party datasets.

For example, a retail organization leveraged external data sources to enhance its workforce-availability analysis and contingency planning. It analyzed epidemiological model predictions and location-specific information—such as whether employees would likely commute to work via city buses, passenger trains, or subways for each zip code where it operates—in conjunction with its internal workforce data on employee segments.

Regional managers were now empowered to more accurately anticipate when and where they'd need to adjust their workforce plans (for example, hire or move associates) and institute contingency measures, such as shortening store hours.

Some people may read this and think, "Sure, that's fine during COVID, but COVID is an extreme anomaly. Do we still need so much access to external data once the pandemic ends?" The reality is, when you are trying to solve almost any analytical problem, there is value—and a significant cost—in adding more layers of external data sources.

For example, if you're trying to predict volume sales for a store and the road the store is on was closed for two to three months, or if a competing shop opened up next door—these crucial factors might be missed by only analyzing the internal data.

Other organizations might put too much emphasis on seasonality and historical performance, which can lead to botched targets or famished or bloated inventory predictions. Poor forecasts can also have dire consequences. Failing to properly identify competition, new retail channels, or a decrease in new stores for distribution can not only hurt your sales but could leave you with more inventory and less income than you expected—or the embarrassment of out-of-stock items.

In this chapter, you'll see how to get the ball rolling with the right external data team and corresponding mindset. Next, you'll delve into Explorium, an automated external data management platform and learn how the platform's unique all-in-one platform gives data scientists, analysts, and business leaders access to all relevant external data signals needed to drive decision making. Finally, you'll see how real customers such as Melio and GlassesUSA.com are using Explorium's unique data enrichments to help them scale up rapidly and instantly prioritize their right potential customers.

Kick Off Your External Data Strategy by Choosing the Right Team

If you want your organization to leverage external data efficaciously at scale, you'll need the right team, the right mindset and strategy, and the best tools. Technology can solve problems and streamline pipelines, but without the right team working together and executing a careful strategy, most projects are likely to fail or never get off the ground in the first place. Everything in business starts with having the right people, so putting together a committed data-sourcing team is paramount.

A recent McKinsey article suggests setting the foundation with a dedicated data scout. A data scout or data strategist must identify operational, cost, and growth improvements that could be powered by external data by partnering with the data-analytics team and business functions. A good data scout should also build enthusiasm for the opportunities that external data offers, pinpoint the best use cases, find the best data sources, and demonstrate the value generation of said data. The data scout should be the quarterback or field general of your external data team—so please choose wisely.

McKinsey recommends other roles that should be drawn from across functions, including:

- Purchasing experts
- Data reviewers
- Architects and DevOps engineers
- Data engineers
- Data scientists and analysts

Purchasing experts steer the ship through murky legal waters and may also be able to connect technology vendors and external data providers, while data reviewers maintain proper compliance with data privacy and other rules and regulations.

Data architects and DevOps engineers, meanwhile, cultivate the relevant infrastructure to support and streamline the use of external data, make sure it's integrated with internal data sources, and manage access to data.

Data engineers must cooperate with both the data science teams and the line of business stakeholders by assisting in the evaluation process of the external data and preparing the data for the data science and analytics teams. Finally, data scientists and analysts apply the external data to their analysis and use cases and quantify the benefits and upgrades in model performance, thanks to external data.

Start with the Low-Hanging Fruit, but Beware of Shiny-New-Toy Syndrome

Now that you have the right team in place, it's time for your external data team to get started. Many teams are often quite enthusiastic, but their zeal can lead them into three common pitfalls:

- Trying to leverage dozens of external data sources without a plan or identified use cases
- Starting with the most difficult use cases
- Diving straight into the implementation of the latest technology without fully considering its suitability for your business, people, or society, or what is known as "shiny-new-toy syndrome"

It's easy to be extremely ambitious and enthusiastic when starting out with external data, and often teams fall into the trap of attempting to leverage a plethora of external data sources at one time without a definitive plan or earmarked use cases.

This common misstep is understandable but may cost you valuable time, money, and resources. Other teams may try to bite off more than they can chew and tackle the most difficult use cases first. Simply reviewing and prioritizing the lowest-effort use cases with the highest positive impact from external data is a more prudent and effective approach.

Racking up a few early wins will instill confidence and earn your team credibility. Your empowered team can then develop a successful pattern that will help them take on more challenging use cases with a higher impact. These more challenging cases will naturally require more effort, but your team will have already been conditioned after picking the low-hanging fruit several times.

Another common trap is shiny-new-toy syndrome. Your team should first identify a clear business problem—augmenting decision making, meeting customer needs better, predicting supply and demand, etc.—and then use the technology to solve the problem.

Start with your North Star and then work backward. It can be fun to buy a new tool to fix up your house, but you don't want to start drilling holes into the walls haphazardly. Get a good builder, identify the problem, get the best tools, and get to work.

Now that you know who needs to be part of the team and have defined the business value and the right use cases, it's time to move onto choosing the right tool—the best external data platform.

Discover Relevant Datasets and Add Significant Value

Melio's mission is straightforward: keep small businesses in business by providing a smart, simple B2B payments solution tailor-made for their needs, such as conveniently paying bills online via bank transfer, debit card, or credit card. The company has been growing rapidly since its launch in May 2019. Monthly active users (MAU) grew by over 2,000% in 2020, and the platform soon faced the challenge of processing billions of dollars in payment volume. The Melio team, however, lacked the capacity to analyze the surge in inbound leads and to identify and authenticate the small and medium-sized businesses (SMBs) that were relevant business prospects.

Melio's marketing team knew they needed a better system to prioritize inbound leads and to validate their eligibility and fit for Melio's products and services. They needed a fast, automated process to identify and prioritize the right business segments and take their mission to scale.

Melio tapped Explorium to support its ambition for hypergrowth by implementing Explorium's external data platform to discover relevant datasets that added significant value. In the end, the partnership resulted in models that broadened lead scoring criteria by using a combination of internal data and external enrichment tools that made a significant impact.

The Melio marketing team is now able to analyze their marketing funnel and identify relevant leads to better focus marketing and sales resources on the most relevant, high-value segments. Melio has been able to achieve the following tangible outcomes:

- Data-driven decisions are based on model outputs, leading to a 15% increase in conversion rates.
- Considerable savings are delivered in the time previously allotted to funnel analysis, which enabled its operations team to achieve a three-fold improvement in efficiency.
- More core tasks are addressed with the same resources.

Explorium's all-in-one platform helped Melio by automatically matching external data with internal enterprise data to uncover thousands of signals to improve ML models and business outcomes.

Other organizations, meanwhile, need help optimizing their digital campaigns. When GlassesUSA.com, the fastest-growing, leading online eyewear retailer, was looking to enhance its Facebook retargeting campaign by pinpointing users with a high likelihood to buy eyewear, they implemented an in-house predictive model they built using the Explorium External Data platform.

GlassesUSA.com built and trained an effective predict-to-buy model with Explorium's automatic machine learning capabilities that assigned a score to website visitors—predicting their likelihood to purchase. Users were bracketed into groups according to their likelihood to buy, and then custom audiences were created on Facebook for each bucket.

After implementing the "predict to buy" model, GlassesUSA.com has seen significant, tangible improvement in the performance of its Facebook dynamic ads campaigns, including:

- A 10% improvement in ROAS (return on ad spend)
- A 10% increase in conversion rates reflected in an increase in the "add-to-cart" ratio
- Enhanced operation and marketing efficiency

Conclusion

From COVID anomalies to footfall traffic to standard marketing, promotion, and advertising campaigns, organizations that can create better infrastructure to collect, store, analyze, and leverage external data—and successfully integrate it into their operations with their internal data—can outperform other companies by unlocking improvements in growth, productivity, and risk management.

Whether you're looking to mitigate risk, predict customer lifetime value, forecast demand, optimize ad spend, or create custom experiences, Explorium can help get the wheels turning by:

- Helping to define your goals
- Enriching your data with augmented data discovery
- Getting your optimal feature set
- Selecting and deploying your model
- Generating valuable insights

Computer scientist Peter Norvig has been quoted many times as saying, "More data beats clever algorithms, but better data beats more data."

About the Author

Joseph D. Stec was born and raised in Cleveland, Ohio (USA) to a family of Polish and Ukrainian immigrants. He studied English literature and accountancy at Miami University before working on the American Stock Exchange in New York City. After witnessing the lingering devastation of September 11th firsthand, he moved to Poland to lecture at university. Joseph has since lived and worked in Ukraine, Russia, Istanbul, Belgrade, Taiwan, and South Korea.

Joseph now lives in Kyiv and is an avid photographer. He has had exhibitions in several countries and he draws upon his photos to support his writing and painting. Currently, he is a freelance writer, working on articles and ebooks related to securities law, big data, and AI. Joseph has finished four books and counting.