EXPLORIUM

Making Sense of Data:
Auditing, Discovery, and Acquisition

Auditing ..... Page 4

Ensure your data is stable and relevant for your predictive challenge

Discovery ..... Page 11

Understand the data you're missing and how to find it quickly

Acquisition ..... Page 13

Fill in the gaps in your data using a step-by-step acquisition process

Table of contents

Understanding Your Data - Auditing and Discovery 3

Step 1: Auditing your data 4

- Where do you get your data? 5
- Auditing your existing data 7
- What data are you missing? 11
- Is your data scalable? 12

Step 2: Filling in the gaps 13

- Why do you need external data? 14
- Filling in the blind spots 14
- External data issues 15
- Data acquisition challenges 17
- Augmented data discovery 18

You've got the data, what's next? 20

Understanding Your Data - Auditing and Discovery

It's easy to get caught up in a rush to start building your machine learning (ML) models. After all, you have data just sitting in your databases, and you've got the technical capital to start working on ML, so why not just dive in, right? You most certainly can, and you might get decent results from it, but you'll be doing yourself and your organization a disservice.

Before you actually build your models, you need to understand your goals and how you plan to achieve them. That process starts with data, which is crucial to ensure that your models will be accurate, relevant, and successful once they reach production. The data workflow begins long before you ever write a single line of code for your machine learning algorithms. After all, before you can get insights from your models, you need the data to train them. The question is, do you have enough data to get the insights you need? And if not, how can you fill the gaps?

In this whitepaper, we'll go over the steps you need to take to audit your data, including making sure it's scalable and relevant. Then, we'll look at why you need external data, how to go about getting it, and some of the challenges you might face while acquiring it.

Step 1:

Auditing your data

ML models are fueled by data. If you want your models to run seamlessly, you need to make sure your data pipelines are reliable and robust — and that the quality of the data you feed in is top-notch.

To start with, this means getting a clear picture of what data you have, how you collect it, where it comes from, and what you're missing.

Where do you get your data?

Every organization already collects data internally. For every data source you use, you need to understand exactly how the data is collected, where it comes from, and how it makes its way into your data storage.

Some examples of data sources you are likely to use in your organization include:

CRM

If you use a Customer Relationship Management system (CRM), you will have a huge repository of personal customer information, demographic data, support and service call notes, warranty information, and logs of interactions. You may also have extensive employee and HR data, sales and purchasing histories, marketing reports, and contact information.

Customer-facing touchpoints

Through your stores and online presence, you are almost certainly harvesting a plethora of data types on your customers, your marketing

performance, and so on. Touchpoints include:

- **Website** - Your website may be tracking user IP addresses for location, details on what people do while they're on your site, where they click and how long they stay, the browsers and devices they're using, and potentially, where they go next or their online habits generally.

- **POS** - For retailers, Point-of-Sale (POS) data includes everything about a transaction and where it takes place. That includes data on your barcode scanners, checkout counters, and cash registers, right up to the shopping area or mall, or the relevant market.

- **Email and direct mail** - With email campaigns, you can track open rates, click-through rates, and conversions, split-test subject lines, and in short, collect data on every aspect of the email, the campaign, and the interaction. Direct mail gives you less oversight, but you no doubt collect data on response rates to different types of campaigns, as well

- **Social media** - Social media platforms like Facebook, Twitter, and Instagram collect mountains of incredibly granular data on user characteristics, interactions, and behaviors. While much of this is kept internal, a significant amount of useful data is packaged for business users or can be integrated with external analytics and machine learning tools.

Auditing your existing data

Whether your data is collected and stored in databases, data lakes, or a data warehouse, there are some fundamental questions you need to ask before you can embark on a data science project.

How is your data organized?

Simply having data is just the first step. Understanding how it's organized can help you better plan for the future. Raw data needs to be properly sorted, harmonized, and stored before it can actually be useful.

Simply having data is just the first step. Understanding how it's organized can help you better plan for the future.

Is your data a mess?

If your data storage is a mess, it will be incredibly difficult to leverage your existing internal data for machine learning. You need to be able to find your way around the datasets you have. You need to make sure that individual datasets and databases have accurate labels so that you can jump to the relevant information and feed the right stuff into your models when you get to that point. You need to be confident that you've used consistent titles and headings for data points so that the machine learning models will know where to make appropriate connections.

When we say "messy", we're often talking about really simple errors and inconsistencies, such as how individual people's names are formatted. A name might be listed as the first name, then last name. Or as the first

A name might be listed as the first name, then last name. Or as the first initial, followed by the full surname. With a title beforehand (Mr., Ms., Dr., etc.), or without. Middle names may be included, either in full or as a single initial. Some entries might use extra punctuation, like a period after an initial. Some may have an extra space left by accident when a name was entered into the system manually. The data entry person may not have noticed a spelling error or basic typo or may have used an alternate spelling of a popular name.

The likelihood of all uses of a name being entirely uniform is incredibly unlikely. This is especially true if the system used to enter them allowed them to be typed freely, rather than encouraging a set style or limiting the range of options — for example, by capitalizing all surnames or offering a drop-down menu for titles. The trouble is, for your database, every instance of this name, no matter how tiny the variation, will be treated as a separate entry.

When you multiply this by hundreds, thousands, or millions of data points, you can start to picture how this might skew your models. Now imagine that other fields in your databases are as messy as the names. That sales and pricing information has been entered in a range of

That sales and pricing information has been entered in a range of currencies, or formatted with different numbers of decimal places, for example. When you input this data into machine models, some may be duplicated, some may be misinterpreted, and much of it may simply be rejected as invalid.

This is a disaster for machine learning, and certainly something you need to figure out now, long before you start building a model. Take the time now to assess harmonization problems in your internal datasets and tables, and figure out what needs to happen to it before you can start to do useful work with it.

Take the time now to assess harmonization problems in your internal datasets and tables, and figure out what needs to happen to it before you can start to do useful work with it.

What data are you missing?

Running a careful audit of your data doesn't just tell you what you already have. It also highlights what you don't have. This is a major benefit of auditing that is often overlooked.

As you go through and clean up your databases, pay close attention to the missing details. What can't this data tell you?

> "Over three-fourths of decision makers want to find new external data sources: weather, news, social media, demographic data, census data, and other socioeconomic indicators. They need more: a little local knowledge, a heads-up on what's to come, the inside scoop - anywhere they can find that differentiating nugget."
>
> - Forrester Research

What do you still need to build your models?

Keep in mind that the goal of a data audit is to make sure you have everything you need when you come to build your models. Now that you have an idea of what information you're missing, ask yourself what kind of data you would need to complete the picture and broaden your scope. This will be vital when you come to the next step of the process.

Is your data scalable?

Finally, before you move on to step two, think about how you can improve your current internal data collection methods for the future.

How do you go about collecting data now? How has this created some of the endemic problems you've flagged in this stage? What bottlenecks and challenges are caused by the systems and processes you use to solicit, capture, and store data internally? Could you make simple improvements to your forms, databases, and so on to eliminate some of these inconsistencies?

The key here is scalability. Once you've built your models, you will almost certainly want to scale up, and for that, you'll need more data. How well would your current data streams adapt to that demand? How quickly and easily could you either collect data or convert data from your existing databases, data lakes, and so on into a format that could be fed into your models? The easier you make it to flip the switch now, the more valuable your models will be later on.

Step 2:

Filling in the gaps

Now that you've figured out exactly what data you have (and more importantly, what you're missing), you can begin the painstaking process of filling in the gaps by acquiring and integrating external data.

Why do you need external data?

In many cases, by this point in your workflow, it will be clear that internal data just isn't enough on its own. Perhaps it's too narrow in scope. Perhaps it's not as complete or accurate as you would like. Perhaps a few extra details could really enrich the bare bones of your existing datasets and give you a fuller picture. Whatever the shortcomings of your internal datasets, adding well-chosen external data to the mix will be crucial in your quest for accurate predictions.

Filling in the blind spots

As we've seen, you will inevitably have now realized that the data you have internally has holes in it. Now it's time to do something about that.

Make a list of all the information you need in order to fill in these gaps. Where might you be able to source that data? Who collects or owns it? Can you reach out to those data owners or providers? Is it readily available? Could you look at Google's Database Search, or Kaggle, or other open sources?

External data issues

But wait! Before you dive right into acquiring external data, you must understand the potential challenges that lie ahead. More data does necessarily mean better data, and while there are thousands of online databases and datasets to choose from, they certainly won't all be relevant to your project. Even if they are, they may not be easily accessible.

According to an Explorium survey, most respondents face more than one challenge to data acquisition:

- 5% don't know what they're looking for
- 23% claim it takes too long to find and test new data sources
- 9% say data is too expensive
- 45% note they have problems with all the above

You can't just go out and purchase every dataset that seems remotely interesting and — boom! — your data woes are over. After all, if you've amassed terabytes of data but only end up using a gigabyte of it, that's not good data or a sensible investment: it's wasteful. Not only that, but those bloated datasets will also potentially be confusing for your models to navigate.

Instead, you need to find smart ways to dig through the available data, extracting just the good stuff — the data that will actually help you make the predictions you need.

> More data does necessarily mean better data, and while there are thousands of online databases and datasets to choose from, they certainly won't all be relevant to your project.

Data acquisition challenges

Acquiring external data — specifically, acquiring only the high-quality, relevant data you need — can be an enormous challenge, especially when you don't have easy access to hundreds of pre-vetted, compatible, external sources.

In particular, you need to think about how you will:

- **Track down** only the most relevant data sources in an ocean of existing data.
- **Budget** for any datasets you will need to pay to access.
- **Test** the datasets you're thinking of using, to make sure they're right for the models you are building. Unless you have tools or platforms designed for this very purpose, it will be tricky to manage without actually buying or downloading the full dataset in advance.
- **Vet** these datasets for quality, making sure they fulfill all legal requirements surrounding data protection, collection, and storage. Again, when you're using datasets from all over the world, it can be hard to keep track of country-specific laws

and regulations, ensuring you don't inadvertently break the rules by accessing datasets that don't meet changing standards, or by moving that data to servers in another location.

- Keep a lid on the **time** your team is investing to acquire this new data. In many cases, doing this manually means it will never be truly scalable.
- Ensure that the data you choose **stays relevant** over time.

Augmented data discovery

Underpinning all these data acquisition challenges is scalability. Trawling through hundreds of data sources, one after another, to locate the parts that are relevant to your models takes a lot of time — especially if you need to go through a legal process and sign an NDA for each one. Few organizations have the luxury of spending months on end just hunting for a single data source and getting it to a place where it can be used.

This also means you can jump straight to the data within the sources that are actually relevant to the models you're building; someone else

has already completed all these fiddly, complicated steps, so you really just need to point and click. Which in turn means, of course, that you can scale up your data discovery quickly and easily, without needing huge investments of time or resources.

> "More successful companies recognize the value in doing that: 66% of firms with revenue growth of 10% or more prioritize external data sourcing, while only 51% of those with slow growth (or no growth) do. Speaks for itself, no?"
>
> - Forrester Research

You've got the data, what's next?

So you've checked your data, and found it slightly wanting. You've gotten the external data you need to fill in the gaps. Now you're done with data, right? Not quite. This is just the start and, while you have the right data for your models, it's still not quite ready to give you the insights you need. If you simply feed it into your models now, you'll be sorely disappointed at the results you get.

Now's not the time to rest on your laurels. It's time to start building a dataset that will maximize your ML's impact and give you the uplift your organization is looking for. The question is, how can you do it? The answer is, it's time to start cleaning, prepping, and getting your data ready for data science. It's time to read on to chapter two of our series, **Making Sense of Data Prep: ETL, Wrangling, Data Enrichment**, and see how to get your data prepared to give you the best results when you start building ML models.

Ready to continue to part two? Head over to Making Sense of Data Prep: ETL, Wrangling, Data Enrichment now.

EXPLORIUM

About Explorium

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

For more information, visit www.explorium.ai