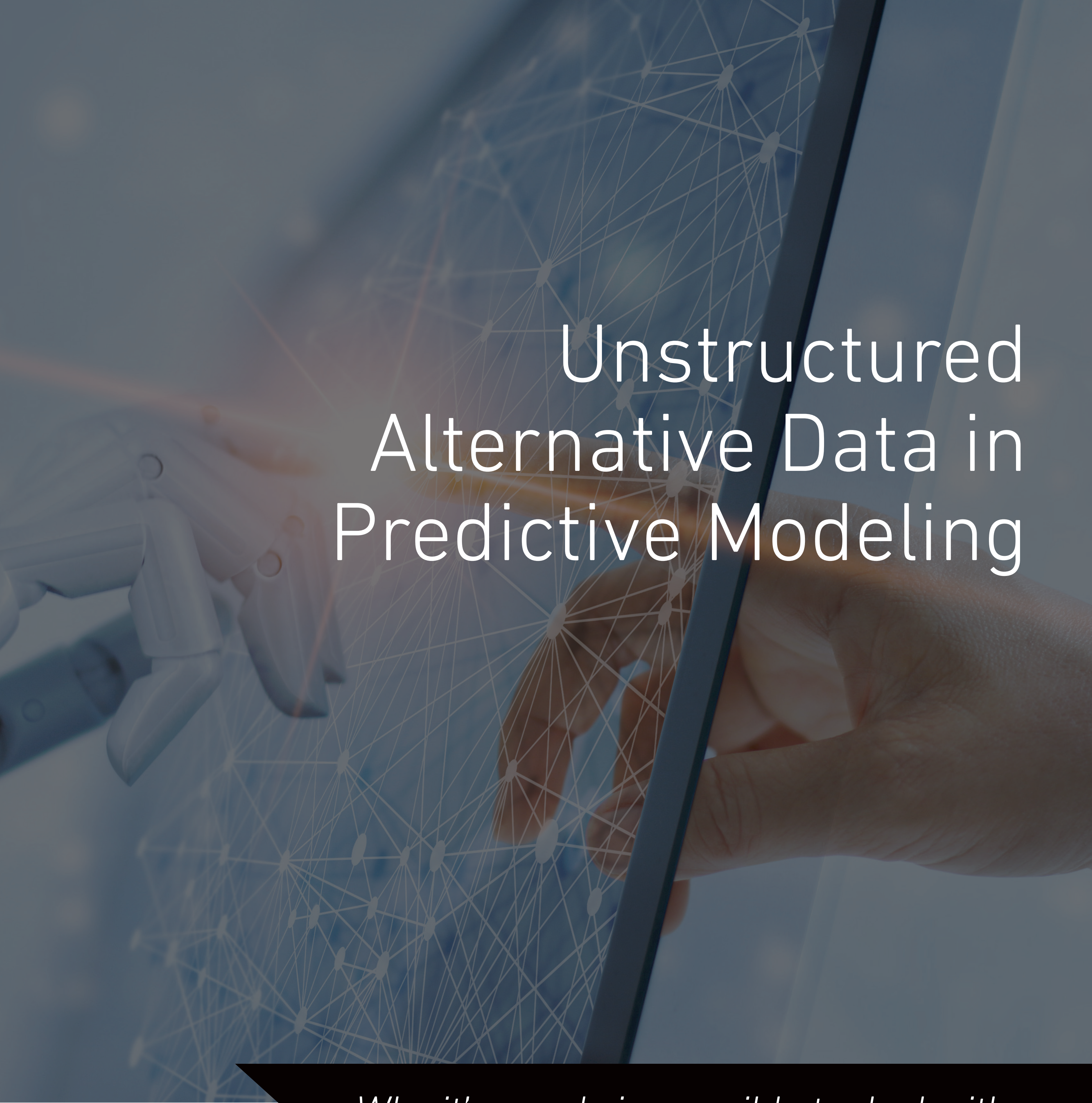


Unstructured Alternative Data in Predictive Modeling

Why it's nearly impossible to deal with but more than worth the trouble.

BITVORE

By Greg Bolcer
Chief Data Officer, Bitvore

What’s alt-data, and why is it important?

Alternative Data (a.k.a. “alt-data”) is any rapidly changing, big data related to business performance, investment, and finance. Alt-data supplements more traditional financial data sources like stock prices or basic company information in financial decision-making processes.

This data can be as simple as tabularizing the sentiment of a company in the news over time or as complex as extracting information from overlooked data hiding among the weeds of day-to-day business activities. This type of data is rarely used and seldom correlated to the performance of a company.

Historically, alt-data comes from paper receipts or various work documents that are not available in electronic format. Other sources include private company information that either is not shared beyond the individual line-of-business or not captured at all. The latter typically happens when the storage costs can’t be justified by any value that the data might provide, even though the data might be valuable when combined with other things.

Alternative data can also be derived from individual or aggregate data by algorithms or machine learning on traditional sources, so that results can be used as inputs to other analyses. These sources can come from news, government agencies, the companies themselves, or from licensing or purchasing data from third-party aggregators.

One especially critical thing about alt-data is that it has a network effect. Network effect occurs when the value of the whole network is exponentially higher than the sum of any individual piece. A telephone, for instance, is worthless if you are the only one in the world that possesses one. The more people that have phones, the number of potential calls between them goes up exponentially, and so does the total value. If there are ‘n’ telephones in the world, the value is **n^2**, which represents the ability for every single phone potentially being able to call every single other one.

Data works the same way. Integrating and correlating more datasets creates more value. Not all combinations of data are interesting, but there are nearly infinite ways you can assemble datasets into collections. Each collection can then also interact with others in valuable ways. The potential value is proportional to **2^n**, which represents all the different ways of integrating the various groupings.

There are barriers to using alt-data. Storing alt-data just for the sake of storing data has a cost associated with it. Most are either expensive to collect/extract or expensive to store and analyze because of the size. While the theoretical value of combining data can be calculated, adopting it into use sometimes takes a little longer due to having to overcome initial cost justifications.

Take web servers, for instance. If you went to your company executives in the early 90s and told them that you were going to use their expensive network connection, clog it up with traffic, run a piece of software on a costly company machine, and allow those outside of the organization to grab proprietary company information your competitors could use against you-the executives would probably fire you.

But that’s exactly what happened with web servers (minus the firing). After the benefits of having a web server far outweighed any initial costs and concerns, companies were able to leverage the collective value. Web servers are a perfect example of a network effect. Alt-data is just starting to provide enough value to overcome the initial costs and concerns, and its adoption will only accelerate from here.

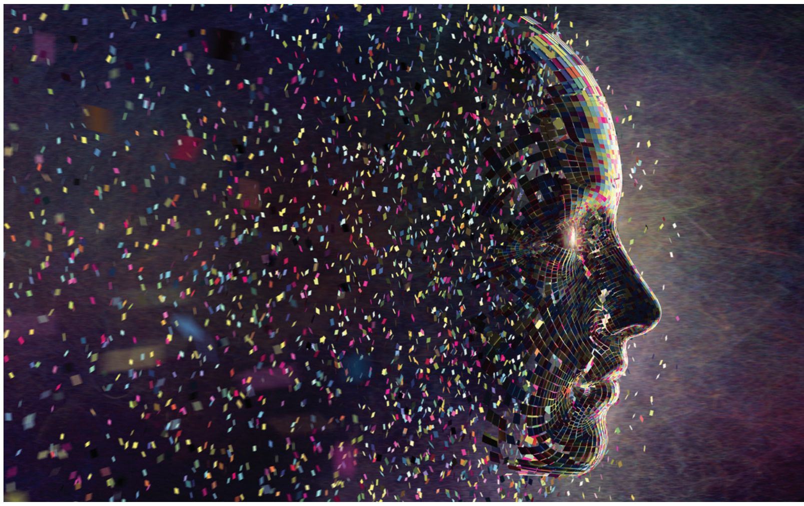
What is unstructured alternative data?

Raw text is considered unstructured data, but the truth is, even raw text comes with some points of structure. What source did it come from? When was it published? Who is the author? At Bitvore, we focus mostly on semi-structured data like a textual news item, though we do look at press releases, SEC filings, investor presentations, public records, ratings, social media, product reviews, job postings, and other information.

Some companies do use more visual alt-data like satellite images of how many cars are sitting in a storage lot or how much foot traffic goes through various airports, buildings, malls, or public spaces. That sort of information, while useful, falls outside of our interest and customer areas.

We can reason and derive structure out of alt-data. Did it come from a reputable source or is the source blacklisted? Did a human write this, or is it robonews/junk? What is the subject of the story, and where did it take place? A lot of these initial answers help us to separate invaluable, valuable, and useless information and can be derived structurally even before we apply more powerful machine learning algorithms.

Another source of semi-structured data comes from websites. The reason websites are semi-structured is that you aren't just looking up values on the site to answer questions. Who is the CEO? Who is on the board? What is the last big deal the company did? For how much? With which customer? When did they last launch a product?



There are web scraping technologies available, but without doing a bit of analysis, it's hard to figure out the information or answers you need. The critical question is, how do you get a machine to understand and answer these questions to the same level of quality as a human sitting down and digging through the website to find the answers? The answer is: humans and machines aren't perfect, but a little machine learning goes a long way in being able to do far more, far faster, and on far more sites than feasible for any amount of humans.

How do data scientists use alternative data to build predictive models for analysts?

There's an urban legend that gets passed along among alt-data data scientists. It starts like an old joke. Two guys walk into a bar. A stock analyst following Tesla is drinking away his sorrows as his clients keep asking him what is happening with Tesla. They keep promising tens of thousands of cars, but every time he visits the company, they are stockpiling thousands of vehicles that aren't moving anywhere.

His friend who works in satellites tells him he can look at the past months' satellite feed as Moffett Field is right across the bay and his satellite flies right over there. It expanded from there to the point people started live-streaming all the distribution centers as a way to try to predict whether there will be sufficient demand for the new model, and ultimately whether the share price will fall or rise.

This excellent example of unstructured data is simply a picture of how many cars are sitting on any given lot at any given time. Some users were even able to write automated counters and live-stream the locations so that traders could have the information on-demand and any time they wanted. The problem with the whole thing is that the alt-data lacked context. As Tesla ramped up production, so did their temporary storage. Without knowing the other factors, having access to the fastest, most accurate alt-data in real-time can be open to any number of wide interpretations.

How does Bitvore's use of alt-data take a different approach?

Alt-data isn't valuable without correlating it to more traditional data sources. The single most valuable source is timestamped news. While there are a lot of things that can be discovered that never show up in the news, having access to those things lacks context without validation in the news. That's not to say all news sources are equivalent. There is a production cycle and an escalation process for specific items. Bitvore is really good at identifying early news items that will be significant before more traditional, slow-moving media covers them.

This expertise helps in predictive models. In the short term, we can find valuable news items by correlating the information with our alt-data and leveraging our machine learning models that have been tuned using tens or hundreds of millions of records across various companies and industries. For longer-term predictions, we look for patterns in our analysis. We identify individual items with something called a signal. A signal is simply an indicator that something business impactful happened with a very high degree of reliability. We also correlate a signal to the company that is mentioned. When we combine both the company and the signal, we come up with precision news, a highly reliable indicator that something significant happened.

Our latest predictive efforts use highly reliable information to predict other signals. For instance, in our muni bond product, if a city eliminates fire, police, or an ambulance service, forgoes teacher raises in a school district, or starts discussing pension costs (all signals in our system), we can predict with a very high degree of certainty that they will be announcing a budget shortfall at the end of the fiscal year. Likewise, if a city announces a budget shortfall, raises new money through issuing new bonds, pushes through public employee raises, or raises property taxes (also all signals in our system) we can predict a city or a county bankruptcy.

Companies follow similar patterns. Fundraising, an abundance of new product launches, executive churn, and various other patterns of signals can result in looking for new money/fundraising, trying to sell the company/merger & acquisition, financial distress, or even bankruptcy. While these types of predictions are not absolute, knowing there is a higher percentage chance over the course of the next few quarters that one of these business impacting signals will happen is invaluable information.

Why do data scientists spend 60-80% of their time dealing with unstructured alt-data?

In short:

- Multiple, disparate sources of data
- Normalization issues
- Cleansing issues

For data science, there is always a tradeoff between using a small, but immaculate data set versus using a large and dirty one. There are many ways data can be muddy. The first is the concordance, or if there are several different names of companies, i.e., Family Dollar Stores, Dollar Tree, Dollar General, Dollar Express, Dollar Holdings. Which company names are the same verses different? Which are still around? Sometimes it's even hard for humans to know. Geographically, we have to differentiate between the City of West, Texas and West Texas, Central Pennsylvania, and the City of Center, Pennsylvania, and hundreds of other ambiguous items.

Likewise, if our computers get a news item from a local newspaper, there are no indicators in the story about which U.S. state the paper resides. We would need a human to go to the site and try to figure out where exactly the story was from by hand. Luckily we've worked out automated solutions for a lot of these issues.

Other things like stemming, a technique where the root word is the same independent tense or usage, cause issues. For example, hospital versus hospitality. The hospitality suite at the hospital was very inhospitable. When trying to differentiate between Marriott Suites, a Marriott-run senior center & hospital, and a public-funded hospital that gives family grants to stay at the local Marriott, it takes a lot of disambiguation.

Even before we get to such issues, there are general cleansing issues. For instance, the headline: Salesforce Signs Definitive Agreement to Buy Tableau. The body of the article is: "Registration required," "no content found," "To continue reading subscribe below," or the widespread "Shop now in our online pharmacy." Typically we blacklist almost a billion individual news records a year. Blacklisting by itself, is practically a full-time job.

Malware, pharmacy scams, paid promotions, celebrities, obituaries, pornography, lifestyles, weddings, coupons, and other non-business related information- and that's just the first pass. In addition to pattern-based blacklisting, we have machine learning models trained to remove old content, error pages, paywalls, dark webs, robonews, and a variety of other things. Interestingly, the single most significant factor in identifying a fake or promotional news story is simply that there is no current date and/or no data comparable to when the story first appeared.

Once the records are cleaned up, we can look at duplicates and similar articles. We keep copies and similar stories in our system for analysis, but other than identifying them, they hardly get analyzed in our system. Once the data is distilled, it can be analyzed for entities/entity-extraction, sentiment, signals, geography, and any other features or data items. As mentioned, mapping those things into like-entities requires careful processing and a lot of machine learning and human expertise.

What is AI-Ready Data?

Definition: Aggregated from multiple sources, normalized to appropriate domains, and cleansed of garbage.

At this point, you have only clean data and reliable entity-extractions, the minimum needed for AI-Ready Data. Adding other clean data values like geographies, signals, sentiment, scoring, or other differentiators allows data scientists to carve off just the data they want.

A large part of that is being able to sort items by a value or only find items that frequently appear together. When you are looking at tens of thousands, or even millions of things, being able to perform extensive data operations to get precisely what you need to do is essential.

Microsoft Excel, one of the favorite tools of data scientists, has a hard limit of 1 million rows. Imagine trying to read a 5 million row data file into Excel just so you can sort, rank, score, and excerpt the top 500,000 things you need for your experiment. For simple filtering, data scientists end up either putting the data into a database or files, or writing scripts to find patterns. Consider the difference between:

- "Paris, Texas", 1.05, FinancialHealth
- Paris/TX, 1.05, FinancialHealth

Sometimes just having a comma in your dataset is problematic. Likewise, when dealing with unstructured text where you need the title of a news article, binary characters, double or single quotes, punctuation, and a variety of other things like character encodings can mess up the best-laid tools.

Joining data with other reference-able datasets is a black art in itself. Imagine you have a record that is a news article about Salesforce. You want to join the information with Salesforce's number of employees. Instead of having a column of data that says Org1's employee count, Org2's employee count, OrgCombined's employee count, you want to be able to do some analytics on the combined employee count by joining the values from some third metadata source.

Other issues include unrolling or grouping. Say you have Salesforce and Tableau in one article, tagged by Bitvore with Mergers & Acquisitions and Financial Health signals. Unrolling lets you figure out when you have two lists of things in two different columns so that you can do better analytics.

- [Salesforce,Tableau],1.05, [MergersBankruptcy,FinancialHealth]
- Unrolled:
  - Salesforce,1.05,MergersBankruptcy
  - Salesforce,1.05,FinancialHealth
  - Tableau,1.05,MergersBankrtupcy
  - Tableau,1.05,FinancialHealth

Both signals belong to both companies. But if you are unrolling a CEO's name and a VP of Marketing's name for a sales agreement, how do you know which company the CEO works at and which the VP of Marketing works at if they are two different companies? Sometimes you need to keep the extracted data together because there is a dependency that shouldn't be unrolled.

Finally, since time is a critical dimension for making predictions, data scientists have to roll up time into hours, days, weeks, months, quarters, years. If you want counts for how many signals for a company happened last month, you will get a number. You can then compare that number to a previous time frame.

Having a strategy and the tools to help solve these issues quickly is what AI-ready data is. Eliminating the 60-80% of time data scientists spend on making data ready for predictive analytics is exactly what Bitvore does. Bitvore creates AI-Ready Data.

Read Greg's latest white paper, Using AI-Processed News Datasets to Perform Predictive Analytics, Monday morning Quarterbacking the Amazon HQ2 Selection.

About Greg Bolcer, CDO Bitvore

Greg is a serial entrepreneur who has founded three angel and VC-funded companies. He's been involved at an early stage or as an advisor to at least half a dozen more. Greg has a PhD and BS in Information and Computer Sciences from UC Irvine and a MS from USC. He started his career at Irvine as a researcher in Web protocols, standards, and applications under a series of DARPA-funded grants. He formerly was the Intel Architecture chair for the Peer to Peer working group and was awarded the Distinguished Alumni of the Year in 2004 from UCI.

Read more from Greg on the Bitvore Blog.

About Bitvore

Bitvore provides precision intelligence derived from world business news and information. Our products are deployed in over sixty of the world's largest financial institutions, allowing them to rapidly create augmented intelligence solutions to address their unique business requirements. Augmented intelligence solutions assist employees in making faster and more effective decisions, so they outperform the competition.

To learn more, visit bitvore.com or contact us at Bitvore.com/contact-us/.