Rethinking Alternative Data in Institutional Investment

ASHBY MONK, MARCEL PRINS, AND DANE ROOK

ASHBY MONK
is executive director of
the Global Projects Center
at Stanford University
in Palo Alto, CA.
amonk@stanford.edu

MARCEL PRINS
is chief operating officer
at APG Asset Management
in Amsterdam,
The Netherlands.
marcel.prins@apg-am.nl

DANE ROOK
is a postdoctoral
researcher in the Global
Projects Center at
Stanford University
in Palo Alto, CA.
rook@stanford.edu

Alternative datasets (alt-datasets) appear to be entering the financial mainstream. Alternative data (alt-data) have always occupied a crucial role in financial markets, but, until recently, cultivation and use of alt-data were largely seen as niche activities for specialist players (e.g., hedge funds with esoteric strategies). Yet, the number and diversity of readily accessible alt-datasets has ballooned in the past decade. This proliferation now confronts institutional investors (Investors)—such as public pension funds, endowments, and sovereign wealth funds—with a dilemma: How can they responsibly choose which alt-datasets are most likely to be sources of significant value for their investment objectives? This article's main goal is to help Investors properly address that question.

Within the financial community, alt-data are widely understood to be datasets that are not conventionally used in investment decision making.[1] A few archetypal (and well-hyped) examples of alt-data have emerged in recent years. These include

- satellite imagery of commercial or economic activity (e.g., the number of cars in parking lots of major retailers, ships passing through ports, and agricultural or mining operations);
- social-media streams, from which consumer, political, or other sentiment may be gauged;
- microdata about consumers' shopping activities (e.g., credit card transactions or in-app purchases on smartphones);
- data scraped from the internet (e.g., job postings to track corporate hiring patterns); and
- data exhaust—the assortment of log files, cookies, and other digital footprints created by people's online browsing (including geolocation data from searches on mobile devices).

These diverse examples are united by a common value proposition for alt-data: market participants can extract an informational edge from some alt-datasets and use it to beat competitors when identifying trading opportunities.[2] This opportunistic,

[1] Some examples of conventional financial datasets include asset prices and trading volumes; corporate earnings reports; economic forecasts of employment, inflation, housing starts, and consumer spending; exchange rates; and yield curves.

[2] Less commonly, some Investors are beginning to view various alt-datasets as sources of insight for responsible investing (e.g., as providing information about environmental, social, or governance impacts of investable companies). As we discuss later in this article, the value proposition of such uses for alt-data does not rely on speed. Nevertheless, such applications largely remain viewed as (at best) secondary applications for alt-datasets by most market players currently active in the alt-data space (although some experts expect it to become more primary over the coming years).

speed-centric perspective on alt-data's value is pervasive and neatly captured by the tagline of a leading alt-data platform operator: "Alternative data is untapped alpha."[3]

We argue that alt-data's core value proposition is, however, meaningfully different for Investors than that slogan would suggest. Investors (as defined earlier) have a distinct comparative advantage over other market participants: patience. Because of their long operating horizons, Investors can pursue investment strategies unavailable to other market players. This comparative advantage is more aligned with defensive and defensible approaches to alt-data than it is with the exploitative strategies that short-horizon investors tend to pursue. That is, Investors will likely be better off using alt-data in ways that are unharmed by competition over alt-data (i.e., nonrivalrous) or for activities others cannot easily replicate (i.e., excludable).[4] In rethinking how alt-data will be most valuable to long-term strategies, we contend that Investors must also rethink how they evaluate and characterize alt-data, along with whom they should partner in gaining access to alt-datasets.

Rethinking these three issues could guide Investors in selecting alt-datasets, and strategies for analyzing and acting on them, that better fit with their organizational contexts. We seek to help Investors re-examine how alt-data could best serve their needs and offer recommendations that are informed by both formal empirical findings and our own close interactions with Investors. We also explore examples of how alt-data can be creatively used in defensive or defensible strategies.

Although building capacity around alt-data is strategically valuable in its own right, doing so has the added benefit of promoting innovation. Using alt-data demands (almost by definition) that Investors depart from the status quo in their decision making. As such, thoughtful design of an alt-data program can drive innovation in all aspects of an Investor's business (e.g., creative improvements in processes, people's skill sets,

and technology). Finding partnerships that facilitate, rather than forfeit, opportunities to innovate and learn from alt-data is therefore a key issue we address and one that is likely to materially affect Investors' success (with alt-data and beyond).

The rest of this article is organized as follows. We first make the case that Investors are better off designing their alt-data strategies around defensive and defensible approaches to using alt-data than aiming to use it for alpha-oriented, opportunistic purposes. We provide examples of creative uses of alt-datasets under these strategies. These examples emphasize how alt-data can be used for deeper understanding of risk and generating operational alpha. We then cover why existing systems for characterizing alt-datasets do not fit Investors' needs. We consider a replacement system that could improve the appraisal of alt-datasets in terms of how well their characteristics align with an Investor's specific objectives and capabilities. Next, we distill our empirical findings about Investors' organizational attitudes on, and capacities for, alt-data. Our analysis concludes that Investors will generally need to partner for access to alt-data and to realize efficiencies in organizing and (pre-)processing alt-datasets. We detail the benefits and costs of partnering with different types of entities and remark on how opportunities for innovation may be a core consideration in selecting alt-data partners. We then describe how the growing arms race around alt-data could affect Investors. Finally, we close by summarizing our findings and highlight additional facets of alt-data strategies that Investors might wish to rethink in the future.

RETHINKING ALT-DATA'S VALUE PROPOSITION

Although alt-data have garnered increased attention in recent years, their use in finance is not new. Alt-data have played an integral role in investing ever since humans first began keeping records of trade: They deepen the connections between financial valuations and real-world sources of value. For instance, some enterprising merchants in ancient Babylon used measurements of the Euphrates' depth and flow to gain an informational edge in trading various commodities (because they realized that these variables were correlated with market supply) (Lo and Hasanhodzic 2010).

What has recently changed about alt-data's role in finance is its degree of accessibility. Perhaps the most

---

[3] See: https://www.quandl.com/alternative-data.

[4] An example of a nonrivalrous application of alt-data is in screening public equities based on sustainability criteria for the underlying companies. An example of excludable alt-data use would be for due diligence on direct investments in startup companies to which an Investor has privileged access (e.g., a university endowment having first access to funding spinouts from its research laboratories). In this sense, an Investor benefits not from an alt-dataset being excludable but from its own ability to use that data being an excludable (i.e., not easily repeated or imitated) capability.

recited example of alt-data in finance is hedge funds counting cars in retailers' parking lots (which supposedly is a leading indicator of sales performance). In the past, such counts had to be made manually, with analysts physically located in or near cars they tracked. Apart from a small number of well-resourced hedge funds, few financial organizations could devote sufficient resources to such a narrow endeavor. Currently, however, these data are accessible through a subscription service to any investment organization inclined to purchase it (thanks to lower costs of satellite imagery).

More generally, the number and diversity of alt-data sources that are readily accessible to financial entities has mushroomed. The tally of large-scale alt-data vendors who specifically cater to investment organizations has gone from a few dozen to several hundred in less than half a decade.[5] The total alt-data sources potentially relevant to investment decision making that can be cheaply and easily accessed is in the many millions. Furthermore, tools for acquiring and processing these plentiful datasets are increasingly user friendly.[6] Alt-data are steadily becoming mainstream.

As a result, the rate at which any one type or source of alt-data becomes conventional—and therefore ceases to be alt-data—is likely to increase. If the value of alt-data is premised on their conferring advantages in faster exploitation of trading opportunities (as is the case for many financial-market participants), then this means the value of any given alt-dataset will probably deteriorate at an accelerating rate because both alt-data and their value are relatively determined. Notice that data may qualify as alternative at any of three levels: the firm, the industry, and the financial ecosystem as a whole. For example, a dataset may be unconventional for a given hedge fund, but not for other funds in the hedge-fund industry. Likewise, some data may be conventional for a given firm, yet be unconventional for most organizations in the wider financial system. When enough organizations make use of any alt-dataset, it stops being alternative at a system-wide level.

Similarly, two relative dimensions help determine the value of any alt-dataset: *rivalry* and *excludability*.[7] Rivalry is the extent to which one entity's use of a resource diminishes its value for another entity.[8] Excludability is the degree to which one entity can prevent another from using a resource. When alt-data's value is premised on allowing market players to better exploit trading opportunities, then alt-datasets will tend to exhibit high rivalry. Moreover, rising accessibility of many alt-datasets is tending to lead to lower excludability.[9] These trends suggest the shelf lives for alt-datasets may be shortening if their value comes solely from helping to exploit opportunities.[10]

Defensive and Defensible Value

When an alt-dataset's value is premised on it improving a market participant's ability to speedily seize trading opportunities, there is an embedded assumption that the participant will need to act quicker than others to realize that value. This value proposition for alt-data implies that alt-datasets should be more useful for financial organizations with comparative advantages in rapid execution.

Speed is, in general, not a comparative advantage for Investors, and for sound reasons: They are long-lived organizations whose success is mission critical for their beneficiaries. Building an investment strategy around speed can greatly increase the risk of

---

[5] Here, we make a meaningful distinction between providers or sources of alt-data (*point vendors*) and alt-data access providers (*platform vendors*). Later, we discuss why this distinction is relevant. For now, we simply note that the number of alt-data vendors vastly exceeds the number of platforms, and this gap is only likely to widen in the future.

[6] These tools may be standalone (e.g., http://scikit-learn.org/stable/) or part of the suite of offerings from an alt-data platform (i.e., an entity that offers not just alt-datasets but also additional support or tools for working with them).

[7] The dimensions of rivalry and excludability are conventionally used to classify economics goods as private, public, club, or common pool. For such purposes, rivalry and excludability are usually treated as binary categories (i.e., something is either rivalrous or nonrivalrous and excludable or not). We see them here as continuous properties.

[8] Rivalry is a *congestion effect*, which is the opposite of a *network effect* (i.e., a resource's value grows with popularity).

[9] This decreasing excludability may become more prevalent as methods for dataset emulation and replication (e.g., statistically synthesizing better proxy datasets) techniques improve. Likewise, the bigger the market for alt-data becomes, the less incentivized many vendors are likely to be, given that they may be able to maximize revenue by selling their datasets to a wider demand base.

[10] A plausible circularity may exacerbate the shrinking shelve lives of alt-datasets: As the number of alt-datasets grows, more value accrues to those market participants that build alt-data capacity, which makes providing alt-datasets that much more appealing for vendors, who then increase market supply further, and so on.

losing unacceptable amounts of capital. Because most speed-oriented strategies are expensive to implement (e.g., they usually require specialized infrastructure or talent), they are often only efficient to deploy when large amounts of capital can be allocated to them. This risk profile for speed-based investing makes it unpalatable for most long-term Investors to undertake directly. In contrast, many asset-management firms (e.g., hedge funds, active mutual funds, or other organizations that extract management fees) can be relatively short-lived entities (i.e., they may not exist after their founders leave), and their failure would usually be less socioeconomically disastrous than it would be for Investors; thus, their cost of allocating most of their capital to speed-driven strategies is far lower.

Investors are also comparatively disadvantaged in terms of agility. As noted, rising rivalry and declining excludability of many alt-datasets means that market participants who attempt to use alt-data to exploit opportunities must be somewhat flexible to succeed; when some alt-datasets lose value from becoming more conventional, others must be sought. Because alt-datasets are largely heterogeneous, organizations that design investment strategies around them need to be agile. The level of agility required for this purpose would overwhelm the data-management and governance systems of many Investors. Although it can be argued that Investors should strive to improve such systems, in many cases it is more pragmatic to align their use of alt-data with their native strengths.

Perhaps the most powerful comparative strength that Investors have is patience. Their long horizons of operation mean that Investors can reap greater gains than other market participants by being more methodical and disciplined in their investment activities. Accordingly, we assert that the deepest value proposition alt-data has for Investors entails *defensive* and *defensible* strategies.

Defensive strategies prioritize capital preservation and prudent risk-taking over speedily exploiting opportunities. Hence, defensive strategies that incorporate alt-data should be centered on pursuits such as advanced risk analysis and management or improving operating efficiencies. Done correctly, these strategies can substantially decrease the degree of rivalry over an alt-dataset (i.e., one Investor building a defensive strategy around an alt-dataset need not reduce the value to another Investor of doing likewise). Risk management and exclusionary screening in responsible/sustainable investing are quintessential examples of defensively applying alt-data: Alt-data can be an invaluable source of intelligence on environmental, social, governance, and other factors that are germane to responsible/sustainable investment decisions, and use of an alt-dataset for exploring those factors does not necessarily degrade its value for use in the same type of decisions by others.

Defensible alt-data strategies, meanwhile, can help Investors increase the excludability of an alt-dataset by either restricting access to it (e.g., via making it proprietary) or by developing execution capabilities around it that are not replicable by other market participants (e.g., through having privileged access to infrastructure deals via special relationships with local governments).

In this article, we concentrate on defensive alt-data strategies because we believe these are most broadly applicable across various Investor types and circumstances. We cover defensible strategies briefly in the final section of this article, and we reserve a detailed treatment for a companion article.[11] From what we see, the two clearest categories of defensive alt-data strategies for Investors are deeper understanding of risk (to better allocate and manage it) and driving operational alpha.

Understanding Risk

Modern efforts in risk management largely emphasize simplifying risk over deeply comprehending its sources. Put differently, such risk-management paradigms are better at detecting that specific risks have materialized in the past than revealing why they have done so. For example, they may uncover how price movements for a given basket of securities correlate when responding to some event, but they deliver scant insight into why the event transpired in the first place. For market participants that operate over short horizons, knowing the correlation may suffice for managing risk, but for Investors to better leverage their capacity for patience, understanding the reasons why can be essential.

This need to more deeply probe causality is due to the fact that correlations in conventional datasets often break down over longer horizons and typically do not reflect the entire spectrum of events that could occur over long periods of time. Alt-data can (partly) mitigate these shortcomings by supplying more context about how events in the wider world drive downside moves

---

[11] See Monk, Prins, and Rook (2018).

in markets. Although it is true that rapid detection of such events might allow Investors to exploit opportunities, a less rivalrous (and more durable) benefit of early detection is that it allows more time for Investors to respond to downside events once they are flagged as likely. Moreover, added context can help warn about unprecedented downside events. When more variables are tracked, there is a higher likelihood of catching anomalous behavior that heralds highly atypical events, even if the precise impacts of such events might not be immediately apparent.[12] The ability to be alerted about unusual events is of prime importance to Investors. Large market crashes practically never play out in the same ways their predecessors did, but a single crash can fully nullify many years of outstanding performance.[13]

The purpose of defensive alt-data strategies is not to totally eliminate risk exposure for Investors but more to distribute it selectively.[14] Selective risk exposure is the chief idea behind smart-beta investment strategies, which seek to control exposures by holding positions in assets that are not necessarily proportional to their respective market capitalizations. Today, many Investors pursue smart-beta investing through purposed exchange-traded funds (ETFs), but smart-beta ETFs often lack fine control over risk exposure. For one, such ETFs are usually only ever composed of public securities and thus are not helpful for controlling private-asset exposure. Second, the asset weightings for the vast majority of ETFs are based on factors derived from conventional data (e.g., company size, dividends, or price momentum). These factors mostly fail to reflect risk in any nuanced way. For finer control over risk exposures through smart-beta ETFs, Investors must purchase shares in niche ETFs that can have high liquidity risk and management fees. Finally, the programmatic rebalancing rules for passive (and many semiactive) smart-beta ETFs

can create unintended—and severely disadvantageous—consequences when abrupt market downturns occur.

Judicious use of alt-data may allow Investors to deploy smart-beta (or similar) strategies in ways that avoid these shortfalls. A suitable supply of alt-data could allow Investors to design index-construction methods for public (or private) assets that create tailored, controlled risk exposures.[15]

The use of alt-data to more deeply understand risk is not confined to portfolio construction. Indeed, alt-data have applications in other areas of risk management, such as in asset oversight and due-diligence processes, especially in private markets. For example, if an Investor directly owns a real-estate development project in an emerging market, it may hire a local manager to oversee that asset's construction. However, this delegation can generate agency problems, such as when the Investor must rely primarily on the local manager's reports about the project's progress. A form of alt-data that might lessen such problems is images of shadow lengths from the project's construction site (e.g., taken from aircraft or satellites). Algorithms such as those developed by Orbital Insight are capable of converting the lengths in such images into calculations of the pace of projects so that an Investor might enjoy greater clarity about whether its local manager is providing valid reports.[16]

An example of alt-data's use for deeper understanding of risk in due diligence involves the analysis of a venture capitalist's networks in determining whether to invest in one of its funds. The relevant networks might be derived from alt-data sources, such as LinkedIn (for general partners' professional and social networks), or

---

[12] Consider a parable example: An island civilization that never has witnessed (or even heard of) a tsunami may nonetheless get advanced warning of an impending anomalous event because of the sudden, dramatic recession of shoreline that characteristically precedes a tsunami.

[13] Long-lived entities are more likely to encounter such crashes, so being able to not do too badly during these crashes is as good as, if not better than, exploitation speed. Investors cannot just shut their doors if they do poorly.

[14] That is, by augmenting information sets with alt-data, Investors may reduce unwanted exposures (e.g., to climate change or reputational risk of investee companies) in a more controlled way, while increasing their desired exposures.

[15] In practice, such methods might be similar to those used by Kensho Technologies to construct its *New Economy Indices*, which capture public companies' degrees of involvement in thematic technological trends, such as artificial intelligence, autonomous vehicles, or drones. To derive its indexes, Kensho uses natural-language processing to identify a company's exposure to a given trend by parsing its public filings (e.g., 10-Ks, 20-Fs) for information on (for example) product lines, supply chains, or planned capital expenditures. Although such filings do not qualify as alt-data, this approach could be applied on other, less-conventional text documents to construct indexes (e.g., sustainability reports).

[16] Another example that may materialize in the future could involve Investors using internet-of-things data feeds from their investee companies or assets. Such data could be used in risk management, help in monitoring human work patterns and information flow, give greater clarity on microjudgments, and help make valuation more real time.

built from scraping websites or digital newsfeeds (to capture what other funds were co-investors on specific deals). Because relationships are integral to most venture capitalists' success, understanding the strength or weakness of a fund manager's networks can be a crucial variable for deciding whether an Investor should allocate capital to that manager.[17]

Some other examples of how alt-data may be used defensively for understanding risk include the following:

- harvesting dynamic pricing information from online sources to garner a clearer, more real-time picture of inflation (and draw on wider or more targeted sources of pricing information than are usual in generic consumer-price indexes);
- aggregating label information (e.g., nutrition facts, ingredients lists) from food-product companies' offerings to see how they may be vulnerable to shifting dietary trends or new warnings by health agencies (Investors may then be able to compel company managers to alter their offerings—e.g., through shareholder activism for publicly traded companies);
- assembling online price and ratings histories of possible competitors (e.g., from Airbnb, TripAdvisor, or Yelp) or price series of airfares to that locale when doing due diligence on candidate direct investments in leisure-related properties (e.g., hotels or casinos);
- using microsensors (or other remote sensors) to track fluctuations in soil moisture for determining what plants are best suited to intercropping in a plantation-forestry investment; and
- controlling reputational risk from investee companies by monitoring controversies about them that arise in social-media posts (or other localized or unconventional news outlets).

Generating Operational Alpha

Alongside deeper understanding of risk, Investors can also use alt-datasets in defensive ways by turning them into sources of *operational alpha*. The chief idea behind operational alpha is to better align operating resources with investment strategies by eliminating internal inefficiencies in how investment processes are executed. This concept is (loosely) related to investment alpha, which is the generation of returns in excess of some benchmark, after adjusting for the riskiness of the assets used to generate the excess returns. Although operational alpha has a secondary benefit of (potentially) improving gross investment returns, its chief aim is to improve net returns by reducing unneeded operating costs. Because such reductions are often risk free, operating alpha can substitute for, and in many instances is superior to, investment alpha.[18] It can also complement investment alpha because it frees up room in the risk budget and thus allows pursuit of strategies with higher upside.

Alt-data can aid Investors in driving operational alpha. Perhaps surprisingly, most Investors already possess large volumes of alt-data within their own organizations. Because alt-data are defined as data not conventionally used in decision making, novel forms of internal data count as alt-data.

Aggregation and disaggregation are key to converting conventional internal data into alt-data. For instance, inventive collation and synthesis of documents (e.g., e-mails, investment memos, and contracts) can uncover precious metadata that is able to provide insights for enhancing communication, culture, negotiation, time allocation, benchmarking, and diligence. Likewise, the disaggregation of collective processes into individual contributions can give a clearer picture of where latent organizational resources—and opportunities to improve them—reside. For example, by tracking how individual internal users query and access documents in organizational databases, an Investor can construct a map of intraorganizational knowledge flows and examine the typical approaches its analysts use in problem solving. More granular visibility of these individual activities can not only expose areas for improvement but also help better identify best practices.[19]

---

[17] More specifically, an Investor may have little ex ante clarity about the specific startup companies in which a venture capitalist will invest (and no control over how it does so once capital is pledged). The quality of the venture capitalist's likely co-investors, however, may be easier to discern and serve as an indicator of the ultimate riskiness of its portfolio.

[18] Notably, operational alpha can be (almost or fully) market agnostic.

[19] Such added visibility of internal processes also has a potential risk-management benefit in the form of compliance. Newly legislated requirements for data handling (e.g., the European Union's General Data Protection Regulation) mandate that users be made aware of how their personal data are being treated. In the case of

Implications of a Changed Value Proposition

In rethinking the value proposition of alt-data, Investors will need to re-examine other views and approaches they have regarding alt-data. Specifically, in pursuing defensive or defensible alt-data strategies, Investors will likely need to alter how they characterize and access alt-datasets. In the next two sections of this article, we discuss pragmatic paths for addressing each of these matters.

RETHINKING HOW ALT-DATA IS CHARACTERIZED

Because the number and diversity of alt-datasets is enormous, Investors need to be discriminating when selecting which alt-datasets deserve resources (e.g., money to acquire; time to store, prepare, and analyze; and capacity to be governed). Such selectivity requires characterizing alt-datasets to establish which will be most valuable for organizational needs. As the value any dataset has to an Investor lies in the questions it can help answer, there is a need for data-characterization methods that can reflect the question-answering capabilities of datasets (alternative or otherwise).

Alt-data are defined in an exclusionary way—by stating what they are not (conventional). However, unlike alt-data's definition, a characterization system for alt-data should not be constructed around exclusion: It is more reasonable to characterize an alt-dataset by those properties that it verifiably exhibits, rather than those it does not. Problematically, however, few Investors—or, for that matter, financial organizations in general—have any such system for alt-data characterization. In fact (and as we will detail later), Investors rarely have any formal criteria for establishing whether a dataset is indeed alternative (i.e., a threshold that divides conventional from unconventional data on the basis of scarcity, novelty, or another relevant quantitative or qualitative dimension).[20]

Unsurprisingly, because few Investors have any systems for distinguishing or characterizing alt-data, few use any consistent process for valuing its worth in advancing organizational objectives. Undoubtedly, rigorous valuation of alt-data (or any data, for that matter) is a difficult undertaking and subject to wide error margins.[21] Characterization is a more achievable step: It at least facilitates judgments about whether a given alt-dataset aligns with organizational capabilities and strategic priorities. Lack of characterization systems, however, invites the expenditure of resources on alt-datasets that do not fit with organizational priorities and resources and promotes avoidable waste.

Apart from being wasteful, not having characterization systems can challenge an Investor's fulfillment of its fiduciary duties or regulatory compliance: Investors may be hard-pressed to claim that they are engaging in responsible decision making when decisions are made based on data that are not well understood (e.g., in terms of blind spots it may create). Suitably understanding data (whether alternative or conventional) in any consistent way requires a means of characterizing it.

Existing Characterization Systems

Existing systems for characterizing alt-datasets are not suitably aligned with the value propositions we have described. These existing systems either ignore the ways in which an alt-dataset is likely to create value for an Investor (and so neglect organizational context) or assume that any dataset's main use will be driving investment alpha (or a similar short-term, opportunistic pursuit).

For example, Kolanovic and Krishnamachari (2017) posited a characterization system for alt-data that focuses on the origins of datasets (Exhibit 1). This system is not ideal for Investors' purposes for several reasons. First, although it encompasses many sources of alt-datasets, it is not necessarily exhaustive. Second, it gives no indication of how valuable a given alt-dataset is to an Investor. Taxonomical schemas such as this are not best suited to help Investors evaluate alt-data.[22]

Investors, these users can be their employees. Because the definition of what constitutes personal data is evolving, Investors stand better chances of remaining compliant if they already have developed processes and systems for tracking diverse forms of internal data in their organization.

[20] More generally, many Investors have no formalized models or system for characterizing data or judging data quality.

[21] Inarguably, an alt-dataset's value should be positively related to its quality. Yet no quality metrics exist that are universally applicable across datasets or free of restrictive assumptions. We must resort to using properties of data that can serve as context-appropriate proxies for quality. It is on these properties that alt-data should be characterized.

[22] Taxonomical systems are characterization systems that are (or attempt to be) mutually exclusive and collectively exhaustive—that is, the items they characterize must fit into one, and only one, classification category within the system.

Kolanovic and Krishnamachari's Characterization System for Alt-Data

Source Category	Specific Alt-Data Source
Individual Processes	Social media, news and reviews, web searches, personal data
Business Processes	Transaction data, corporate data, government agency data
Sensors	Satellites, geolocation, other sensors

Source: Kolanovic and Krishnamachari (2017).

Kolanovic and Krishnamachari (2017) proposed another taxonomical schema for alt-data characterization that does embed a value proposition and strives to indicate the usefulness of alt-datasets in relation to use cases based on asset class and investing style. Unfortunately, that system is premised on investment-alpha generation, and so it does not cover defensive or defensible uses, which thus undercuts its relevance for Investors (which is further lowered by being taxonomical).

Dannemiller and Kataria (2017) avoided the taxonomical approach and instead suggested that alt-data be characterized on a "continuum … from structured to unstructured." For the purposes of indicating the likely value of an alt-dataset, using continuums, and not discrete categories, makes sense, but whether a dataset is structured or unstructured does not immediately reflect its value for an Investor. It is true that more effort may be required to extract insight from unstructured datasets (which makes them more expensive from an organizational-resource perspective), but this does not necessarily reflect the full value that an alt-dataset holds. For example, both unstructured and structured alt-datasets may be relevant (or not) for defensive or defensible approaches by Investors.

Although big data and alt-data are not perfectly identical, there are cases in which alt-data qualify as big data. It may thus be hoped that characterization schemas for big data could sometimes be applicable to alt-data. The most prevalent such schema is the *3 Vs* of big data: volume, velocity, and variety. IBM's Big Data unit suggests a further dimension: *veracity* (i.e., the degree of uncertainty around a dataset).[23] These systems are a step

in the right direction because veracity, velocity (the rate at which new data arrive), and volume (the size of a dataset) could all potentially add to a dataset's value for an Investor.[24] Yet these dimensions by themselves are incomplete, and none seem to squarely encapsulate how specific properties of an alt-dataset should translate into value. For example, velocity may be important for assets that have value-determining properties, which change frequently, but not so important for those without such properties (e.g., many private assets).[25] Thus, freshness—how well a dataset reflects the most recent changes that are material for decision making—might be more appropriate. Likewise, volume seems to be less important for Investors than whether a dataset is comprehensive. That is, a dataset may contain many items (i.e., have high volume) from only a narrow number of categories of interest. In such a case, a dataset that has smaller volume, but encompasses more categories (i.e., is more comprehensive), would likely have higher value. We thus need a different characterization scheme.

The system devised by Kitchin (2015) comes closest to what Investors need. It builds upon the 3-Vs setup (but is still intended for characterizing big data, rather than alt-data) by adding four additional dimensions: *comprehensiveness*, *granularity* (how fine- or coarse-scaled the data are), *relationality* (how many fields a dataset shares with other datasets of interest), and *flexibility* (how easily new fields can be added to a dataset).[26] Comprehensiveness and granularity seem to be apt fits for Investors' purposes, but it is less clear that relationality or flexibility are pertinent concerns. Furthermore, Kitchin's scheme gives no explicit consideration to the known quality (i.e., reliability) of data. Knowing how reliable a dataset is can be essential for Investors to decide how it can be used.

Six Dimensions of Alt-Data

We adapt Kitchin's (2015) system by replacing relationality, flexibility, variety, and volume with the dimensions of *reliability*, *actionability*, and *scarcity* (and replacing the velocity dimension with the more fitting notion of *freshness*). Reliability (which covers the

---

[23] See: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[24] Velocity may concern the rate at which new datasets are onboarded or the rate at which existing ones are refreshed.

[25] Velocity may also be valuable (for example) in rapidly detecting reputational risks for Investors in social-media activity.

[26] Kitchin actually uses "exhaustivity" and "resolution" in place of comprehensiveness and granularity, respectively.

Exhibit 2
Six-Dimensional Characterization of Alt-Data

Dimension	Explanation
Reliability	How accurate, precise, and verifiable the data are (e.g., error-free, unbiased, checkable)
Granularity	The scale covered by specific data points or entries (e.g., continental, industry-wide)
Freshness	Age of the data (i.e., when collected/generated) relative to the phenomena they reflect
Comprehensiveness	What portion of a given domain the data cover (e.g., 25% of households in Canada)
Actionability	Degree to which significant actions or decisions can be made based on the data
Scarcity	How widely or readily available the data are to other (especially competing) organizations

Source: Authors.

accuracy, precision, and verifiability of a dataset) seems to us a more fitting concept than IBM's veracity. Reliability essentially equates with the known quality of a dataset.[27] Actionability and scarcity are loosely related to, but distinct from, the ideas of rivalry and excludability. In a sense, actionability and scarcity are primitives of rivalry and scarcity. For rivalry to matter, an alt-dataset must be actionable (i.e., it needs to be usable for decisions that lead to actions). Likewise, when rivalry is a concern, it is valuable to have access to scarce (albeit relevant) datasets. Excludability refers to scarcity that is (semi-)permanent. Hence, this characterization schema helps clarify not only what kinds of questions can be answered by a particular alt-dataset but also what kinds of strategies that an alt-dataset may usefully inform (see Exhibit 2 for further details on each of these dimensions).

How do these six characterization dimensions meaningfully contribute to an alt-dataset's potential value in defensive or defensible strategies? The first three dimensions' contributions are relatively clear-cut (although they are also relevant for opportunistic strategies). Because alt-data's purpose is to guide decisions, it should be trustworthy (and, in some cases, transparently verifiable). Similarly, decisions can be made at different levels, and those made at highly specific levels often require very fine data, whereas high-level decisions can usually be made on less granular (or, at least, more highly condensed) data. Lastly, decisions should not be made on stale data for which more recent versions exist. High freshness is thereby desirable in most cases.[28] What qualifies as *high*, however, can vary with the nature of the decisions that are made based on the dataset in question.

The chief way our proposed characterization is more applicable to defensive and defensible alt-data strategies than it is to opportunistic strategies is in importance of comprehensiveness.[29] For opportunistic uses, alt-data need not be comprehensive: They can encompass narrow ranges of instances or categories and still deliver genuine advantages. Although narrow alt-data can still be useful for defensive or defensible purposes, comprehensive datasets are generally more valuable because they give more complete visibility and scope. This greater breadth of coverage is useful for a deeper understanding of risk situations or internal inefficiencies (for defensive strategies), as well as for more exhaustive awareness of ways in which defensible advantages might be vulnerable.

Actionability is highly important for both defensive and defensible approaches because alt-data that

---

[27] Reliability includes how verifiable a dataset is. Verifiability here has two aspects: (1) how readily a dataset's accuracy can be confirmed by using other datasets and (2) the clarity of its provenance. Also note that the first four elements (reliability, granularity, freshness, and comprehensiveness) may be seen as referring to a dataset's richness. Importantly, these four dimensions appear to be the most objective and universally applicable across Investors (it can be argued that scarcity depends on substitutability, which may differ for some Investors—depending on their specific organizational contexts): These dimensions could therefore potentially be standardized to some degree to allow faster assessment of alt-datasets. This might be a useful enterprise for some commercial organization (e.g., an alt-data platform vendor) to undertake in the near future (it also is one that could generate considerable efficiency gains for Investors).

[28] Desirability of low latency does not mean longer time series of alt-data are less valuable. Latency in the case of time series refers to the most recent record in a series. The length of the time series instead reflects its comprehensiveness.

[29] Although we expect that this characterization will likely be useful for many financial-market participants, we realize that the relative importance of each dimension will likely differ across entities (or different types of financial entity).

cannot be translated into proactive or reactive actions are of little (or no) practical value to any Investor. Scarcity has different bearings for defensible and defensive strategies. For the former, its value is more directly connected to excludability. For the latter, scarcity is more related to the rate at which alt-data spread to different financial organizations. If some alt-dataset is very accessible (e.g., public information) and many organizations begin noticing and acting on it at once, there can be systematic effects, which can be troublesome from a risk-management standpoint. Meanwhile, alt-datasets whose scarcity declines slowly can enable more considered and advantageous reaction.

RETHINKING ACCESS TO ALT-DATA

In addition to rethinking the value proposition of alt-data and how they are characterized, Investors might need to rethink how they access alt-data. Indeed, the first two reconsiderations are irrelevant if Investors cannot access alt-data. How any Investor should appropriately access alt-data is a joint function of (1) what entities can provide it and how they go about doing so and (2) what the Investor's current organizational capabilities in and attitudes toward alt-data are. Answers to these questions will necessarily vary to some degree across Investors. Our research indicates, however, that some generalizations can be made so that a typical recommendation can be safely made to Investors. Succinctly, we find evidence that Investors are eager to tap the potential benefits of alt-data but are, on average, not (yet) adequately equipped to independently source, process, and maintain alternative-data resources. However, these current circumstances do not suggest that Investors should abandon efforts to build internal alt-data capabilities by surrendering all alt-data functions to third parties—especially to external asset managers. Instead, we find it reasonable that Investors should prioritize partnerships with platform providers of alt-data (at least for the near-term future).

In the remainder of this section, we first explore empirical evidence on Investors' current capabilities in, and organizational stances on, alt-data. We then turn to how these findings intersect with the different alt-data access modes available to Investors. A focal component of our analysis here is how alt-data can be used as an accelerant for various forms of organizational innovation.

Empirical Findings on Alt-Data in Institutional Investment

The findings reported in this subsection are drawn from extensive interviews with senior decision makers across a diverse sample of institutional–investment organizations, along with results from a survey of Investors. We describe these studies more extensively later, but we first give an overall synopsis.

Succinctly, Investors' current relationships with the rise of alt-data can be described as considerably interested yet significantly underprepared. More fully, we observe the following:

- Investors pervasively believe that alt-data can be used to improve net investment returns, but many are unconvinced that their organization is well equipped to use alt-data to do so.
- Few Investors have a formalized strategy regarding alt-data or are actively developing one.
- Many Investors worry about alt-data costs, specifically to develop in-house capability.
- Investors widely view building or acquiring proprietary alt-datasets as a way to succeed with alt-data and feel that the most valuable use of alt-data is in identifying opportunities.

Both survey evidence and content from interviews provide rationale for, and additional details on, these summary findings. Regarding the former, our survey instrument was completed in February 2018 by senior decision makers (i.e., chief executive officer, chief information officer, chief technology officer) from 22 leading institutional–investment organizations. Collectively, respondent organizations manage over US$1 trillion; they represent a diverse mix of geographies (Australasia, Europe, Middle East, and North America), fund types (sovereign wealth funds, endowments, public pension funds), and fund sizes.

Although 70% of respondents feel that alt-data could help improve risk-adjusted returns in their organization, 90% state that their organization has no "defined alternative-data strategy" (of the 10% that do have alt-data strategies, all admitted that these strategies are "not well developed"). Furthermore, less than 15% claim their organization is "equipped to handle" multiple forms of alt-data (nearly 30% strongly disagree that they are equipped). Less than one-third report that alt-data

are a "priority" for senior management, although 60% of respondents note that their organization is actively monitoring developments in alt-data or considering creating capacity in alt-data.

In aggregate, these response patterns depict an uneasy tension. Investors are clearly aware of alt-data's potential benefits but are not situating themselves strategically to reap these benefits. This awareness-without-progress could drive a reactive need to catch up in the future and cause alt-data strategies to be less carefully designed than they might have been with proactive planning.

Respondents also believe speed and quality are significantly more important properties for alt-data than are granularity or volume.[30] Over 80% claim "opportunity identification" to be the capability that alt-data could improve most within their organizations ("risk management" was selected by less than 10% of respondents). These answers indicate a view that the primary beneficial application of alt-data is in allowing rapid detection of mispriced assets (e.g., arbitrages).

Finally, among survey respondents, a lack of "suitable ways to invest" (i.e., actionability) is stated to be the "biggest challenge" to their effective use of alt-data (32%), followed by the state of their existing technology (23%), analytic capability (23%), organizational culture (18%), and trust in alt-data from key decision makers (4%). We comment on the gravity of these challenges shortly.

To validate our survey findings and probe the situations behind them, we conducted a series of in-depth, semistructured interviews with seven of the respondents (one-third of the full sample). Interviews were conducted by telephone and lasted between 30 and 45 minutes. Overall, these interviews not only confirmed results from the survey but also provided additional details germane to Investors' perspectives about alt-data. First, none of the interviewee organizations have formal definitions for what constitutes alt-data. Such definitions are, arguably, a prerequisite for prudent alt-data strategies. Second, interviewees voiced concern over both the costliness of acquiring alt-data and their organizations' ability to be competitive in their usage of alt-data. Worries about cost fixate on how expensive interviewees

think it will be to conduct alt-data operations in-house. Relatedly, although respondents generally feel that they could become as capable as their peers in developing alt-data functions, they are unsure about whether they can compete with other entities (especially hedge funds) when it comes to their ability to use (i.e., analyze) alt-datasets. Third, interviewees confirmed the survey finding that rapid identification of mispriced assets is the application for alt-datasets with which Investors are most (and, for some, exclusively) familiar. Fourth, a consensus emerged among interviewees that alt-data are most valuable if they are proprietary.

Two other notable points arose in the interviews, concerning (1) data provenance and (2) cooperation. For some Investors, a key stall point is how transparent an alt-dataset's lineage is (i.e., how clear is knowledge of its source, what transformations have been performed on it, and who performed them). Several interviewees noted that their organizations would have reservations about making decisions based on alt-data of uncertain provenance and that murky provenance could dissuade or prevent them from using third-party alt-data. Furthermore, most interviewees agree that their organizations would very likely cooperate with peers in building alt-data capacity.

Modes for Accessing Alt-Data's Benefits

In sum, the preceding observations strongly demonstrate that Investors do not appear prepared to go it alone in sourcing, processing, or maintaining either a wide or deep array of alt-data. Yet, the results also indicate that Investors seem sufficiently interested in alt-data to be unlikely to ignore it altogether. Nor should they. For reasons already mentioned, alt-data could serve Investors as a crucial resource. The question then surfaces of how Investors should access alt-data.

We see two assisted paths Investors might follow in accessing alt-data. The first involves trusting external third parties (including asset managers) to provide indirect access. That is, those access providers take care of the difficult tasks of sourcing, managing, and acting on alt-data, and Investors reap some of the benefits that they may have otherwise received from handling the alt-data themselves. This path addresses the realization that accessing alt-data should not be an end goal in its own right for Investors. Instead, they should aim to maximize the benefits from alt-data.

---

[30] The fact that respondents do not feel alt-dataset size is of primary importance is reinforced by the fact that a majority (72%) answered that alt-data are not "essentially the same as big data."

Nonetheless, offloading alt-data responsibilities onto access providers deprives Investors of a pivotal benefit that building alt-data capacity could provide to them: accelerating innovation. Investors generally struggle with innovation (Monk and Rook 2018). Alt-data, however, supply a springboard for innovation. By definition, the use of alt-data in decision making requires at least some innovation by Investors. In many cases, the amount of innovation itself may be modest, but the amount of learning from it (which could drive future innovation) can be significant.

Moreover, alt-data is a topic that invites considerable excitement and stirs imaginations: It is a sexy concept in finance. Investors can often struggle with innovation simply because they lack internal agreement (within their organizations) about what resources deserve innovation. Alt-data's allure could make it a common point of agreement for coalescing support for innovation.

As we elaborate later, outsourcing alt-data capabilities—such as relying exclusively upon external third parties for indirect access to alt-data—could cause a sizable sacrifice in innovation capabilities for Investors. We believe that many, if not most, Investors should be thinking about how to build in-house capacity around alt-data, especially for defensive and defensible strategies.[31] The degree and nature of this capacity will need to vary with each Investor's own organizational context, but every Investor is indeed capable of building such capacity—to at least a minor extent.

The drive to build some internal alt-data capacity—coupled with the fact that Investors are not ready, by and large, to undertake the sourcing and management of alt-data all by themselves—suggests the second assisted path by which Investors may feasibly access alt-data: alt-data vendors. Two main types of alt-data vendors can be distinguished: point vendors, who offer either a single or limited number and type of alt-dataset; and platform vendors, who tend to offer wider selections of alt-datasets and may additionally offer integration or analytical tools that aid use of alt-datasets.

In the following, we compare prospects and demerits of Investors seeking alt-dataset access through both kinds of vendor, in relation to one another and in relation to external access providers. On the latter, we focus on the impacts of Investors relying on external asset managers for alt-data.

External Asset Managers as Access Providers

Some external asset managers (e.g., some hedge funds) have enjoyed relatively lengthy experience in working with alt-data—at least when compared to Investors. Given Investors' widespread desire to gain exposure to the benefits of alt-data but lack of full capacity to do so at present, it may seem advisable that they seek indirect access through such managers. If doing so came only at the cost of forfeiting some experience with learning to innovate, this option might be recommendable. However, there are at least three additional reasons why it is not. The first stems from the opportunistic nature of most external asset managers. In general, external managers are less incentivized to be concerned about capital preservation and are more motivated to fixate upon investment alpha than are Investors. These differences are not by themselves inherently problematic, given that external managers often are able to build stronger comparative advantages in generating investment alpha than are many Investors (although such advantages are routinely on a gross basis and may not hold once costs are fully considered). What is troublesome, however, is the fact that this emphasis on alt-data for opportunity identification and exploitation predisposes external asset managers to becoming engulfed in an escalating arms race around alt-data. We discuss the drivers, dynamics, and likely implications for Investors of that arms race in the next section.

A second major reason why it might not be recommendable for Investors to rely too heavily on external managers for alt-data access involves transparency and provenance. When Investors outsource their alt-data efforts to external managers, they lose the ability to inspect, verify, and otherwise work with the data on which those managers are basing decisions. Not only does this loss translate into opportunity costs from forgone innovation opportunities, it also creates issues around lack of visibility and verifiability. In not directly accessing alt-data used by their external managers, Investors are forced to rely on those managers to establish and maintain their quality. As we explain in the next section, however, heightening competition over

---

[31] If Investors are electing not to build in-house capacity, then we recommend that the decision result from thorough analysis of long-term trade-offs to the organization (e.g., from loss in potential innovation versus resource absorption).

alt-data may well push external managers to accept and execute investment decisions on alt-datasets of increasingly lower quality, which can inject unforeseen (and sometimes unidentifiable) risk into Investors' portfolios. The likelihood of transparency problems will probably worsen as competition over alt-data grows; managers should then tend to be more secretive about their processes around and sources of alt-data.

Another major reason why Investors should restrict reliance on external asset managers in accessing alt-data is the subsidization of a capability gap. That is, although some external managers presently possess some comparative advantages over Investors when it comes to alt-data, those advantages need not be permanent. Whenever an Investor contracts an external manager to invest on its behalf, and the manager makes use of alt-data to do so, that Investor is effectively subsidizing the manager in improving its capacity for alt-data relative to the Investor's own capacity. This subsidization thus increases both the manager's comparative advantage and the Investor's reliance on external parties for alt-data capacity, reducing the Investor's future strategic flexibility around alt-data.

Access through Alternative-Data Vendors

The path of building increasing internal capacity around alt-data through partnering with vendors mitigates or eliminates many of the aforementioned problems with relying on external managers. First, vendors are (usually) just providers of alt-data, the use of which is determined by Investors. Hence, vendor-supplied alt-data do not necessarily expose Investors to problems connected with opportunistic usage of alt-data (although, as mentioned earlier, many vendors do stress the alpha-generating merits of their datasets). Second, concerns about transparency are partly lessened when Investors access alt-data directly through vendors rather than indirectly through external managers; in the former instance, Investors are actually able to examine the alt-datasets. To be clear, being able to actually work with the data directly does not eliminate the possibility of errors or other quality problems in the data. Yet such possibilities are typically more investigable (i.e., Investors may be able to request assurances about the secure provenance of the alt-datasets) than they are with external managers. Furthermore, because quality and trustworthiness are dimensions on which vendors compete with

one another, many are incentivized to remain highly transparent.

A third concern that is alleviated by partnering with vendors rather than asset managers is that of subsidization. It is true that whenever an Investor subscribes to or buys an alt-dataset from a third-party vendor, it is subsidizing that vendor's comparative advantage in sourcing alt-data (and possibly cleaning or preprocessing alt-data, depending on the services that vendor provides). When creating defensible strategies around proprietary alt-datasets, this subsidization may be problematic. However, we expect that most Investors will instead favor defensive applications of alt-data, in which case such subsidization would actually tend to be helpful for Investors: It would help fund the vendor's provision of additional alt-datasets, and so would further benefit Investors.

Additionally—depending on its infrastructure and particular method of accessing alt-data from vendors—experimenting with different forms of alt-data may be substantially easier through vendors than through external asset managers. That is, switching between vendor subscriptions is, in many situations, likely to be less arduous than switching allocations to different external asset managers. Thus, partnering with vendors may allow Investors to try out more configurations of alt-data when attempting to incorporate it into their strategies, thus increasing their odds of finding a good fit.

Still, the path of accessing alt-data via vendors is not without its downside.[32] The foremost of these is the low degree of excludability for vendor-supplied alt-data. Of course, when Investors' use cases for alt-data are predominantly defensive, excludability becomes less worrying. Likewise, when Investors use alt-data to build capabilities that are defensible (even when the alt-data upon which they are based are not), such as privileged access to deal flows, excludability is not a concern.

Moreover, higher (if not total) excludability can often be achieved at higher cost: Vendors may be willing to provide more exclusive access to alt-datasets for premium prices. In many cases, therefore, Investors that

---

[32] One particular challenge that Investors may face in relying on platform vendors to access alt-data is whether external data provided by the vendor can be easily integrated with the Investor's internal data—without Investors losing control over their internal data or giving others access to it. Tackling this challenge could help vendors distinguish themselves.

access alt-data through vendors can balance dataset cost against scarcity. Striking such a balance may frequently entail working with multiple vendors. In so doing, any Investor should consider the relative advantages and disadvantages of point and platform vendors.

Point vendors tend to be more specialized than platform vendors.[33] The former therefore can often provide more novel, differentiated alt-datasets. Moreover, because point vendors have fewer product offerings than platform vendors, they may be able to verify a larger fraction of their data more intensively than platform vendors (although this need not always be true). Point vendors, however, often have smaller markets for their offerings than do platform vendors, which can bundle together multiple datasets to broaden their appeal. This narrower market for many point vendors means that their costs can be higher than their platform counterparts and so put them out of reach for smaller Investors (or those with less budgetary room for alt-data). Also, point vendors can face diseconomies of scope that are less severe for platform vendors. For instance, it is typical that platform vendors can deliver alt-datasets in a single format or offer more streamlined integration (through, e.g., standardized APIs).[34] Doing so simplifies access for Investors—relative to having to integrate multiple, distinct formats from point vendors.[35]

We anticipate that many Investors who partner with third-party vendors to serve their alt-data ambitions will select a limited (e.g., one or two) number of platform partners and supplement the alt-datasets offered by these platform vendors with specific alt-datasets accessed through point vendors.

---

[33] Examples of platform providers that specialize in alt-data include Neudata and Quandl. More traditional financial-data platforms, such as Bloomberg and FactSet, also are increasing their alt-data offerings. Interestingly, a new type of alt-data entity also seems to be emerging that offers analysis of specific types of alt-datasets, rather than just providing access to them (in some instances such entities do not provide access to the alt-datasets themselves). Examples of these new kinds of entity include Orbital Insight (for satellite-image analysis) and Predata (for social-media analytics).

[34] Integration difficulties may (initially) favor platform providers that specialize in conventional data but offer alt-datasets as an additional service. Increasingly, incumbent providers of conventional data are also offering alt-data.

[35] Platforms may also prove a more efficient way for Investors to keep pace with changing data regulations, under the assumption that the chosen platform can be trusted to stay current with data legislation and related compliance issues.

THE ESCALATING ALTERNATIVE-DATA ARMS RACE

Rethinking alt-data—in terms of its value proposition, characterization, and access—will almost surely be a strenuous process for most Investors. Might it not be better for some to avoid involvement in alt-data altogether? We think not. As we explain here, an arms race around alt-data is underway and gathering momentum across financial markets. The ways in and extent to which we foresee this race escalating lead us to believe that Investors will not be able to escape becoming meaningfully affected by it. We advise that they try to proactively engage with alt-data by building defensive and defensible alt-data strategies, rather than being dragged along in a reactive manner.

Arms-Race Logic

In an elegant application of formal economic logic, Grossman and Stiglitz (1980) proved that the persistence of efficient equilibria is impossible in financial markets. They did so by highlighting a fundamental paradox. Market efficiency is driven by profit-motivated market participants who aim to exploit the mispricing of financial assets through transacting, based on information they possess. In transacting, they jointly increase market efficiency and decrease the value of their information. In (the strongest forms of) equilibrium, there is no unexploited information and so no incentives for participants to either transact or seek out additional information to exploit. However, because the wider world is never in stasis—new information is arriving all the time—markets cannot be permanently in equilibrium. If they were, then there would be no (nonrandom) transacting, which would permit existence of unexploited information and thereby mean that no equilibrium existed, by definition.

Paradoxically, competition is a force that makes markets more efficient but also ensures that they cannot become entirely efficient. An unceasing inflow of new information and data is the key to this seeming contradiction. If no new data about the wider world were to be created, then markets would (hypothetically) settle into equilibrium, but because the world is ever changing, there is continual production of new information.[36]

---

[36] More than 90% of all digital data that have ever existed was created in the last two years (see, e.g., Henke, Libarikian, and Wiseman 2016).

Ongoing competition among market participants to exploit this new information and data squarely qualifies as an *arms race*, which is definable as a situation in which parties are locked in perpetual efforts to outcompete one another, without a defined endpoint. Thus, any effort at active investing amounts to participating in a data arms race. Still, this race is useful: If all participants were passive, then markets would not function.

Every Investor is therefore directly affected by active investing, even if its own strategies are fully passive. By merely deploying capital in public markets (which every Investor does), they are exposed to the active-investing activities of other parties, which affect the volatility and liquidity of their own portfolios. Much of this active investing is done by non-Investor asset managers, who are either hired by or compete with Investors. Thus, all Investors are directly affected by the arms race for data in general (not just alt-data) that is continually underway in public markets. To better understand the consequences of an alt-data arms race for Investors, we should understand what drives the intensity of data arms races more broadly. To that end, rivalry and excludability are core forces.

Role of Rivalry and Excludability

The intensity of data arms races is fueled by the rivalry and excludability of the datasets based on which their participants aim to make investment decisions. Practically all data in finance are rivalrous in the sense that any use of data for transacting reduces (or even eliminates) the value in executing similar transactions thereafter, regardless of who conducts them. This property means any (profitable) actionability of data is eventually self-eliminating so that the value of a dataset decreases by acting on it. This self-eroding value of data's actionability can, however, be partly offset by scarcity. The fewer entities that have access to a dataset, the more proportional value can be kept by those with access. Scarcity is a crucial reason why alt-datasets can be so precious. Most conventional datasets in finance are nonexcludable.[37] Entities with them cannot readily bar others from getting them, and when they transact

on these datasets, others can better divine their content, which devalues them more (i.e., they devalue when first transacted on as a result of decreased actionability, and then again from reduced scarcity).

Alt-data, meanwhile, are typically more excludable—and so any specific alt-dataset tends to be scarcer—than are conventional data. Some alt-datasets can be *permanently excludable*: Those who create or acquire them first can prevent all others from possessing and transacting on them. More typically, alt-datasets are *limitedly excludable*: Entities with them can only exclude others from acquiring them (or replicating them, to some approximation) for a limited time or else can only restrict the number of others who obtain them to a limited extent. Consequently, excludability of many alt-datasets means substantial value can be realized by being first to capture a dataset, even if it cannot be immediately acted on (i.e., scarcity might offset low near-term actionability).[38]

This interplay among competition, rivalry, and excludability underpins the intensity of current land grabs for alt-data (i.e., an alt-data arms race) in global financial markets. Moreover, the combination of these factors creates perverse incentives for market participants to (1) overweight specific facets of alt-datasets when evaluating them, (2) focus on short horizons, and (3) potentially overprice alt-datasets of undetermined value. Valuing an alt-dataset is an uncertain business. Its full richness (i.e., comprehensiveness, reliability, granularity, and freshness) is often hard to establish without spending much time working with it. Likewise, the complete set of ways in which it is actionable may not come into focus until it is more thoroughly processed and analyzed. These layers of uncertainty mean that a hierarchy often emerges for alt-datasets, whereby scarcity and immediate actionability trump other characteristics.

The primacy of these two factors, in light of the limited excludability of many alt-datasets, means

---

[37] This low excludability is increasing the need for financial organizations to conceal their digital activities (i.e., reduce their digital footprints) so that their data and information inputs are less inferable by other, competing organizations.

[38] Alt-data that concern sustainable/responsible investing may be somewhat different from other forms of alt-data in this respect. Investors may well benefit from reducing the excludability of alt-data that are relevant for sustainable/responsible investing (e.g., that relate to environmental, social, or governance factors or sustainable development); in doing so, they might benefit from the emergence of stronger standards and norms regarding sustainable/responsible investment practices.

short timeframes can easily become overemphasized.[39] First, accentuation of datasets that have immediate actionability naturally biases use of them toward the short term. Second, limited excludability creates an impetus to act before others are able to acquire or create substitute datasets. Third, the outsized value of scarcity can encourage *data hoarding*, whereby entities leap before looking and obtain alt-datasets that promise high scarcity and excludability but only minimally consider the actionability of such alt-datasets upfront. Data hoarding can lead to *strategic misfits*, that is, alt-datasets that are poorly aligned with organizational capabilities or priorities and so have low long-term strategic value. Datasets with sufficiently low value can drive pursuit of shorter payback periods to offset their costs and thus compress the time horizons of decisions made with them.

For entities that can cope with, or even excel at, concentrating on short horizons (e.g., some hedge funds), the current intensity of the alt-data arms race may be meaningfully beneficial and increase rewards for their comparative advantages in speed or agility. In general, Investors are not in this group. By and large, their foremost advantage is patience and the ability to operate over long timescales. Unfortunately, Investors' involvement in this arms race is not readily avoidable, which is a problem because the race shows no sign of abating soon. On the supply side, an increasing number of sources and formats for new data continues to emerge. Meanwhile, proliferation of advanced analytic tools, such as deep-learning platforms, are stoking fiercer competition over alt-data.

Sticky Consequences for Investors

Few, if any, Investors will be able to successfully decouple themselves from the alt-data arms race. Its stickiness will mean that Investors cannot insulate themselves from it and still achieve current risk and performance targets. A pivotal realization here is that market competition makes alt-data a moving target. In not using alt-data, market participants handicap themselves by limiting any informational edge that they can possess over other participants. As more participants begin to acquire and transact on any specific type of alt-data (if not the same alt-dataset), however, that type starts to become conventional data, which then lifts the net value of other unconventional datasets. In short, opportunity costs for many market participants, especially non-Investor asset managers, become too great to not seek and use alt-data. As more market participants embrace alt-datasets, markets (especially public markets) will be more affected by them, until they affect even passive investing.

A vital question for Investors engaged in predominantly passive strategies is how alt-data's increased influence over market activity will change the character of that activity itself. How will the rising intensity of the alt-data arms race alter the nature of risk in markets? There is a reasonable case to be made that the increased intensity of this race will not lower volatility in public markets. Indeed, the opposite appears to us more probable, due to (at least) three factors. For one, pressures toward short-termism that we discussed earlier bias decisions toward action rather than inaction. More market activity means greater volatility. Furthermore, intensified competition over alt-data means that there is pressure not only to act fast but also to act big because of fleeting actionability. Possessing a unique and excludable alt-dataset does not block other entities from eroding its actionability by acting on it first: There is reason to act not only swiftly but also extensively to prevent actionability from evaporating. More extensive activity also increases volatility. Finally, the increasing use of algorithmic methods for trading based on alt-data will likely contribute to higher market volatility. Increased volatility will probably raise costs of passive investing through a combination of higher transaction costs (because of faster turnover), hedging costs, liquidity threats, and cash drag.[40] Whether these negative possibilities might push more Investors away from passive strategies is not yet clear.

---

[39] Alt-datasets that are perfectly excludable can still create bias toward short-term action. In contrast, alt-datasets with limited excludability carry additional pressure because of their wasting nature, which can encourage use-it-or-lose-it mentalities. Furthermore, many limitedly excludable alt-datasets are cheaper and quicker to capture than are perfectly excludable ones.

[40] One way to temper risk in passive investing is to increase portfolio allocations to cash, versus the market portfolio. Because the return on cash will not necessarily be increased because of higher volatility in the wider market portfolio, there will be likely be opportunity costs in gross (and possibly net) returns when cash allocations increase (i.e., cash drag).

But if the alt-data arms race succeeds in shifting more capital to active-investment strategies, then a circularity might arise: More money pumped into active investing would raise the value in using alt-data for active investors, which would increase the intensity of the arms race around alt-datasets. This is a perilous treadmill for Investors and threatens their interests.

We have already asserted one way to avoid stepping onto that treadmill: concentrating on cultivating defensive and defensible alt-data strategies. Such approaches could partly immunize Investors against the arms race over alt-data but, by themselves, may not be sufficient. To properly insulate themselves from the alt-data arms race, Investors might need to bolster their capabilities in real-asset investments, such as natural resources and infrastructure. These types of investment have risk profiles distinct from public securities and naturally lend themselves to more defensive applications of alt-data. Moreover, real-asset investments generally allow Investors to more fully exercise their comparative advantages in long-term investing. Rethinking the value proposition for alt-data could therefore go hand in hand with rethinking the composition of long-term portfolios.

SUMMARY

The rising accessibility and diversity of and competition over alt-datasets presents Investors with novel challenges. We believe that these challenges give Investors cause to rethink how they will strategically engage with alt-data. Escalating competition for alt-datasets means that Investors are unlikely to remain unaffected by alt-data and that their strategic planning should take this fact into account. Potential opportunities afforded by defensive and defensible alt-data strategies give Investors ample reasons not only to seek access to alt-datasets but to build internal capacity for working with and acting on them. Cultivating such capacity could be a key engine for innovation.

Although we see many merits for Investors in directly engaging with alt-data, we point out that not all Investors should do so in the same ways or to equal degrees. Defensive and defensible alt-data strategies should be designed in ways that respect the specific resources and organizational contexts of individual Investors, which necessarily means that such strategies will differ from one Investor to the next. However, they need not differ so extensively that Investors cannot beneficially work together in growing their capacities for alt-data, including collaborating to generate and share alt-datasets with one another. We investigate these collaborative opportunities in a companion article.

In closing, we remind Investors of the advantages in being open-minded about alt-data and specifically about taking a wide view on how they can leverage alt-data that already exist in their own organizations. Such data need not be exotic or complicated to be valuable. Indeed, the rising sophistication, but user friendliness, of many data-science tools should cause an increasing number of internal alt-datasets to be significant sources of operational alpha within the immediate future. Moreover, Investors should bear in mind that alt-datasets that relate to internal operations have a very valuable property: They are maximally excludable and thus a fully defensible form of data.

ACKNOWLEDGMENTS

REFERENCES

Dannemiller, D., and R. Kataria. 2017. "Alternative Data for Investment Decisions: Today's Innovation Could Be Tomorrow's Requirement." Deloitte Center for Financial Services, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-dcfs-alternative-data-for-investment-decisions.pdf.

Grossman, S., and J. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70, no. 3 (June): 393–408.

Henke, N., A. Libarikian, and B. Wiseman. 2016. "Straight Talk about Big Data." *McKinsey Quarterly*, October, https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/straight-talk-about-big-data.

Kitchin, R. 2015. "The Opportunities, Challenges and Risks of Big Data for Official Statistics." *Statistical Journal of the IAOS* 31 (3): 471–481.

Kolanovic, M., and R. Krishnamachari. 2017. "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing." Report, J.P. Morgan, May 18.

Lo, A., and J. Hasanhodzic. *The Evolution of Technical Analysis: Financial Prediction from Babylonian Tablets to Bloomberg Terminals.* New York: Bloomberg Press, 2010.

Monk, A., M. Prins, and D. Rook. 2018. "Playing Data Defense in Institutional Investing." Mimeo.

Monk, A., and D. Rook. 2018. "The Technological Investor: Deeper Innovation through Reorientation." SSRN eLibrary, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3134078.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.