Q1 2020

The Future of Data

GREENWICH
DATA | ANALYTICS | INSIGHTS

REFINITIV™
DATA IS JUST
THE BEGINNING

CONTENTS

Ken Monahan is a senior analyst on the Market Structure and Technology team, specializing in electronic trading, regulation and fixed-income markets.

Executive Summary

Advances in the technologies for gathering and utilizing data are progressing rapidly, making it possible to apply new techniques to old problems across society. The financial services industry is among the vanguards of the data revolution. It contains an entire data ecosystem that provides a way to understand the economic and regulatory forces at play. It also provides a window into how datacentric enterprises develop to advance and connect that ecosystem.

The financial services industry is also a pioneer in the technologies driving the data revolution. Financial firms have been pushing the envelope in the development, utilization and commercialization of "alternative data." They are also early adopters of the cloud and advanced analytics such as machine learning and artificial intelligence. These technologies, though advanced, are in their infancy. Understanding the forces propelling and channeling them is essential for discerning their trajectory.

Financial services are also at the forefront of utilizing data for the good of society as a whole, such as fighting organized criminal networks by mobilizing data technologies against money laundering. The industry is also utilizing data to help investors direct their savings toward investments that suit their values. Socially responsible investing has swept the financial system, and data on the environmental, social and governance (ESG) outcomes corporations produce has become fundamental to investor decision making.

The ESG phenomenon illustrates the kinds of questions raised by the advanced utilization of data. Which datasets are most effective? How should the data be gathered? Many of these questions remain open, but it is important to understand the possibilities of data and the context in which they must be answered.

We are at the beginning of an era in which the utilization of data will affect many aspects of our lives. Understanding the forces driving the advance of the data economy is essential for getting the most out of it, in particular for society as a whole. The financial services industry provides an excellent lens through which to view this because it has been at the forefront of the technologies that use data. It has also been quick to apply those technologies to adjacent and socially beneficial uses.

Introduction

Technological progress has become so rapid and so consistent that many simply take it for granted. While technological advances have improved human life, constant marginal improvements can make it difficult to detect changes to the central features of human activity. Such a change is occurring. In every corner of society, new tools are emerging that are changing the way people define, detect and address their challenges. At the heart of this is data.

Two factors are behind this. The first is the nature of data itself—the data economy is driven by scale and the growth in the production and availability of data is geometrically greater than it was just a few years ago. When trying to draw conclusions or make predictions based on data, the larger the dataset or the more datasets to which one has access, the stronger the conclusions and the more accurate the predictions. The technology used to gather, store and process data also has economies of scale. This has drastically reduced the cost of strengthening conclusions and making more accurate predictions.

The second, driven by the first, is that data and the tools with which to utilize it are now accessible on a wider basis than ever before. Analytic techniques that were once available only to the most advanced parts of academia and to a handful of governments are spreading into the everyday world. The cloud, natural language processing (NLP), machine learning (ML), and artificial intelligence (AI) are illuminating possibilities for the utilization of data that are wholly new. These tools are being applied to many issues, from the vitally important like the improvement of regulatory practices and enhancing the quality of investing decisions even, to the everyday, like putting the "smart" in smartphones.

Financial services, the oldest data business, is on the leading edge of this transformation and is an excellent window into data advances due to its unparalleled data demands. It is also a place where questions about the future of data are asked first. What data is available today and what data will there be tomorrow? How can financial institutions harness the power of data to better allocate capital, invest more sustainably or fight financial crime? How will regulators shape the application of these possibilities and how might they use these new tools themselves? Many of these and similar questions will need to be asked by the rest of society as the technology continues to spread.

The financial industry rewards pioneers that embrace statistical methods and data analytics to drive unique methodologies, so it has been an early adopter. For example, 90% of financial services firms are using machine learning in at least one application and 75% of them describe their investments in new data technologies as "significant," according to Greenwich Associates data. When considering the future of data, the financial data ecosystem is the place to begin.

Financial Data Producers and Consumers: Mapping the Ecosystem

Financial Data Producers

Financial data is created, transmitted, transformed, and utilized inside a coherent ecosystem, populated by producers and consumers who are often one and the same. The foundation of this ecosystem is reference data because it provides a locus and a language for the other forms of data. Reference data is produced by groups, generally industry bodies, that determine standard identifiers for different datasets. For example, the Association of National Numbering Agencies (ANNA) is responsible for generating and maintaining international securities identification numbers (ISINs). Among its members is the Committee on Uniform Securities Identification Procedures, which assigns CUSIPs to securities issued in the United States. Some are private companies, such as Refinitiv or Markit, that have legal entity or reference entity databases for determining precise counterparties for transactions or reference credits.

Corporations and governments produce fundamental data about themselves either to better inform their stakeholders or as the result of regulatory requirements. Additionally, banks and independent researchers supplement this self-reported fundamental data with research they produce. The advent of "alternative data" led to the entry of a wide range of new data providers. Some of these are firms that specialize in the creation and commercialization of alternative datasets. Others are firms that have found or believe that they can monetize "exhaust" data, generated in the course of their core business that may have value to firms outside their industry.

TRADITIONAL FINANCIAL DATA

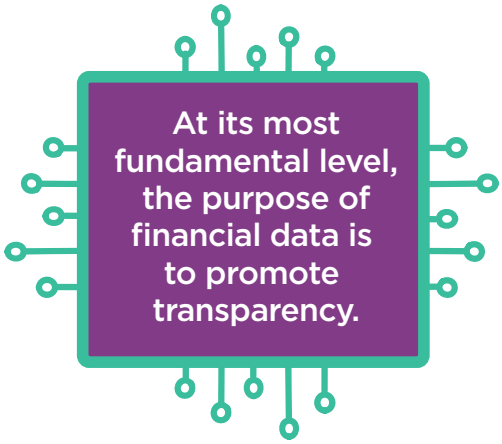


Reference Data *Identifiers that make it clear to what the data refers*	Fundamental Data *Data that describes the underlying conditions of a financial actor*	Market Data *Data related to securities transactions*
ISIN codes	Corporate earnings	Pre-trade—bids or offers to buy or sell a security
Stock symbols	Tax revenues	Post-trade—records of transactions that have occurred
Legal entity identifiers	GDP	Closed-loop accountability

Source: Greenwich Associates 2020

The creators of market data are market participants themselves. Its raw materials are their intentions and actions as represented by their quotes and transactions. In some cases, the participants are responsible for reporting their transactions and so directly create market data. Others require that participants report transactions to a central authority that then releases summary statistics. For example, the Hong Kong Monetary Authority (HKMA) requires participants in the over-the-counter derivatives markets to report their transactions to the Hong Kong Trade Repository, which in turn reports market summary statistics to support transparency to the wider marketplace.

In other markets, such as equities, market participants are brought together by trading venues. These venues take a variety of legal forms: exchanges, multilateral trading facilities, electronic crossing networks, and alternative trading systems, to name a few. Venues have a powerful interest, and often a legal requirement, to gather and disseminate quotes and transactions. This facilitates the process of bringing market participants together. The more quotes and transactions on a venue, the more attractive it is for market participants. In markets where venues dominate trading, some transactions still occur away from the venues, and often the authorities require these to be transmitted to the wider market as well. These services are provided by Approved Publication Arrangements in Europe and by Transaction Reporting Facilities in the U.S.



Financial Data Consumers and Their Use Cases

Financial data is very widely used. Liquidity providers, along with brokerages that provide access to markets, are its most-immediate users, but it has many consumers beyond the financial services firms. Investors require financial data for rational decision-making about where to channel their savings most productively. Corporations also rely on financial market data both to guide them in their business planning as well as in their capital raising decisions. Finally, the public and the regulators who represent them have a strong interest in financial data for what it tells them about the economy generally and about what the future may hold for them.

The whole of society could be said to have an interest in financial data. At its most fundamental level, the purpose of financial data is to promote transparency. Society, and the economy that sustains it, are ultimately built on trust. Trust requires accessible truth, and in the financial system, this is what financial data represents. It is a way for economic actors to ascertain what is true and to make their decisions accordingly. Material disclosures of corporate performance are essential for supporting quality corporate governance. Individual investors with access to financial data can use it to hold both their brokers to their best execution obligations

and the companies to whom they entrust their savings accountable. This has the effect of reducing the power imbalance between the system and individuals: an essential social function.

In addition to this important social function, financial data has its uses within the financial system. Liquidity providers use many forms of financial data when they make markets for their clients. They may hedge their exposures with other instruments and so must know the prices of possible alternatives. Any asset valuation requires a combination of reference, fundamental and market data. Reference data identifies which data elements are relevant to the task; fundamental data is required to get a sense of the underlying business of the security issuer; and market data illuminates what market participants think the future may hold for those fundamentals.

For any given security, multiple sources of reference, fundamental and market data might be relevant. For example, factor investing is an investment strategy based on the theory that security prices can be driven by both macro- and micro-factors and looks to value individual securities formulaically using data about the larger economy, the sector and the individual firm. Thus, the price of any given asset is always related to the prices of any number of other assets, as well as to the fundamental data of the issuing organization and the wider economy. In this way, factor investing represents in miniature how financial markets and their data are interconnected with one another and with the real economy more broadly. What unites them all is data.

What is true for valuing any given instrument or strategy is true of financial decision-making as a whole. The interconnections among different markets and market participants mean that decisions must rely on data from multiple sources. The more data to which one has access, the greater the quality of the decision. This makes access to data from multiple sources, and the integration of that data into a coherent whole, extremely valuable. A class of firms focused entirely on helping their clients navigate the market data ecosystem has arisen to fill this need, and no description of the ecosystem would be complete without them.

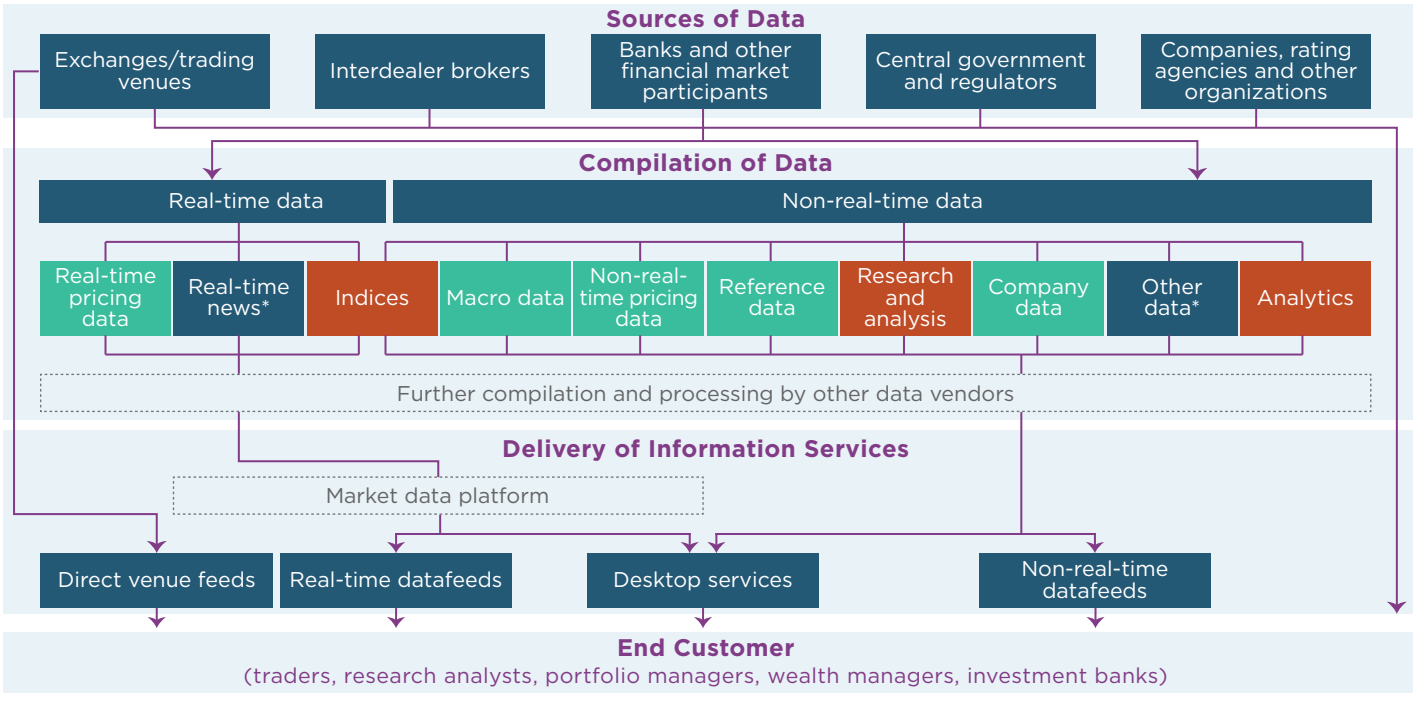
Datacentric Enterprises: Working in the Eye of the Data Storm

The key to rational decision-making is access to data. The more plentiful and varied the data available, the better the decision. This creates a problem for financial market participants: The sources of financial data are as diverse as the data itself. It can be difficult to discern just which sources of data might be useful or most relevant. What's more, data

comes in different formats, at different intervals and at different speeds, presenting technical challenges to collecting it even once the sourcing issues have been resolved. For every market participant to invest the time and the resources necessary to engage each potential source within the financial data ecosystem would be cost prohibitive for all but the largest firms, and even these would find it expensive. This has created a market for datacentric firms that specialize in gathering and distributing financial data.

Datacentric firms create significant value in the seemingly straightforward task of sourcing financial data. They gather data across asset classes, including fixed income, equities, commodities, and foreign exchange, as well as across products, cash and derivatives. Refinitiv, for example, in addition to the other asset classes mentioned, gathers equities data from over 400 venues and fixed-income data from over 800 contributors across 70 venues. But there is more to the sourcing of data than simply locating as many providers as possible. The ownership of data and the rights associated with it can be complicated. A datacentric company that distributes financial data must verify that its data sources have the right to sell it and must themselves secure the rights to the data and secure the right to redistribute it. This step alone creates significant economies of scale for data companies' customers, who might find it wearisome to negotiate bilaterally with so many different sources. It is much easier to have a single relationship that grants access to multiple sources of data.

FINANCIAL INFORMATION PRODUCTS: OVERVIEW



Raw data Derived data * Can be raw or derived

Source: Courtesy of Refinitiv

The data that the data company acquires from these multiple sources will not only include multiple types and asset classes, but the data itself will come in different formats. The variety of formats complicates the efficient consumption of data. It is expensive for each individual data user to translate data in various formats into a standard that would permit the firm to utilize multiple datasets. Instead, data companies convert data from disparate formats into their own standard data model with its own symbology (a form of private reference data). Additionally, data produced at source can be of varying quality. In some cases it may contain errors or format or time stamp differences, and data firms and distributors can detect and rectify these issues centrally before redistributing the data to their clients.

A centralized data firm can convert data from disparate formats into a standard data model, enabling their clients to realize significant economies of scale.

Having this done by a centralized firm enables their clients to realize significant economies of scale. This process is analogous to the container revolution in maritime shipping. Merchant shipping, with a wide variety of cargoes of different shapes and sizes, was onerous and time-consuming, involving a lot of careful positioning to maximize the use of space in the loading and a lot of sorting during unloading. Packaging diverse cargoes into shipping containers of uniform dimensions and labelling them externally with uniform symbology made it easy to optimize space when loading and greatly simplified the sorting and routing process when unloading.

A core business of datacentric companies is distributing data. Getting data to consumers spread all over the world, each of whom has a unique set of requirements, is a significant technical challenge. Not only do different clients need different datasets, but their time horizons can also vary significantly. Some need data within milliseconds of creation, while others can reduce their costs by using delayed data. Some require only the most recent data, while others need historical data for all time.

To serve their clients' diverse needs, datacentric enterprises build and maintain large and sophisticated data networks. The variety of client demands requires flexibility and high levels of throughput because the volumes of financial data can be extremely large. For example, the Refinitiv network delivers 40 billion market updates daily. These networks need to be built with significant excess capacity because the supply and demand for financial data can surge unexpectedly during periods of increased market activity. The Refinitiv network, for instance, can process up to 7 million updates per second across 70 million instruments. By comparison, the consolidated tape for the NYSE and regional exchanges processes 900,000 messages per second. For many users, these networks are mission critical and so require 100% uptime. This, in turn, requires significant redundancy and failover capacity in the event of an outage or fault. In the event of an issue, support staff must be available 24 hours a day on a global basis.

THE REFINITIV NETWORK

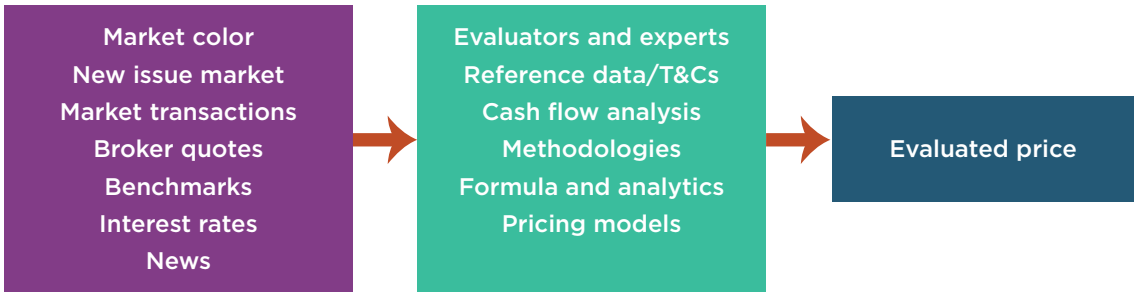
- **40** billion updates daily
- **7** million updates per second across **70** million instruments

Financial data consumers vary in their needs. Therefore, data companies must not only normalize financial data on its way into the network,

they must also ensure that it arrives at the customer in a usable format. For some, this means that the data firms provide not merely the data, but also the applications used to receive and utilize it. The most sophisticated firms take the rawest data and their own applications can use it directly. For others, datacentric firms may offer a terminal or desktop software which will have specialized applications to help clients utilize and visualize their data. Such a firm might also offer a fully electronic data-delivery platform, which transmits data directly to clients who use either their own or third-party applications. Many retail investors get access to financial market data through retail brokerage platforms that, in turn, receive their data from large-scale datacentric enterprises.

Sourcing, normalizing and redistributing data in usable form comprise the core of the data business, but many firms build significant value-added businesses. Data firms develop tools that interact with, enhance and analyze financial data, and help their clients to generate significant insights. The normalization of market data not only makes it easier to distribute, but it also enables the firms to generate metadata (literally data about the data), which simplifies the process of combining seemingly disparate datasets. This has helped reduce the barriers between datasets and enable better pricing and risk management, as well as provide new ways to interpret the effects of news and other events on the financial system as a whole.

INDEPENDENT, TRANSPARENT AND ACCURATE EVALUATIONS



Source: Greenwich Associates 2020

Datacentric companies provide many value-added services, which require a variety of the strengths of market data companies. A good example is “evaluated pricing.” Many securities trade infrequently, which means that valuations based on recent transactions can become stale and not reflect the asset’s true value in the current environment. Evaluated pricing uses information from different sources and the prices of other securities to establish a theoretical price for a particular instrument. This requires access to data from a wide range of sources and sophisticated analytics to weigh and synthesize the data in order to produce a single, realistic price. This is a typical example of how market data firms bring together their access to data, their capacity to normalize it, and their technological and business sophistication to produce value for the client.

Market data companies are required to comply with local regulations and often provide data and services essential for the regulatory compliance of their clients, in particular with regard to their best execution obligations. They also provide a platform where market participants can easily access data whose disclosure is required by the regulatory authorities. While performing these functions, data companies are also market participants competing among themselves. This makes them proactive, working to satisfy their clients' demands.

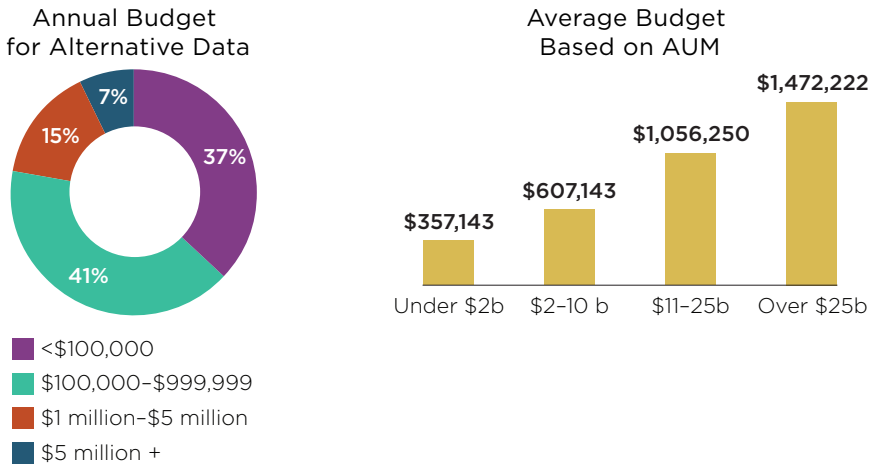
This is what has driven their investments in advanced analytics and in data structures, which permit financial data to be utilized flexibly. They invest in their networks and infrastructure to handle both the increasing volumes of data and the new, unstructured forms it is taking. It is important to note that they make these investments in competition with other firms. This gives them a totally different and more flexible incentive structure than those of mandatory consolidators. Given the speed and complexity of the changes that are underway in the data world, flexibility and competitive innovation are more necessary than ever.

The Future of Data: Alternative Data and Advanced Analytics

The financial data ecosystem is undergoing a remarkable and fundamental change. The historically simple classification of financial data into reference, fundamental and market data is being upended by new data sources that defy categorization almost entirely. This "alternative data" represents a potentially fundamental shift and is driven by two forces. The first is technology. The electronification of many aspects of commerce and society makes it possible to observe and record phenomena that would have once passed unnoticed. This has enabled companies to gather data to a depth and breadth that they never could have before. Some firms are finding that data they gather about their operations, which may have been of marginal use to them, can be extremely valuable to others—and other firms are beginning to specialize in generating this kind of data. The alternative data revolution is the development of a market for these new sources of data.

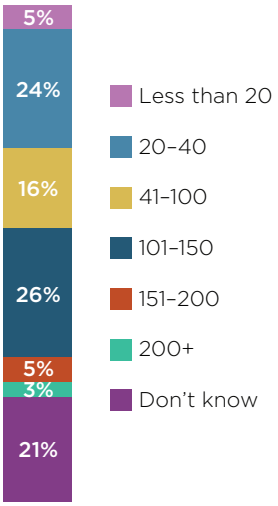
The second force is regulation. Governments have realized that the ease with which data can be collected means that it is increasingly important to ensure that the rights and privacy of citizens are protected. Prudent regulation is necessary to channel development of alternative data into its most productive uses for society, while securing the rights of individual citizens. Thus far, these two forces have worked in tandem to produce a robust and dynamic new marketplace in which significant investments of time and capital are being made.

BUDGET FOR ALTERNATIVE DATA



Note: Based on 27 respondents.
Source: Greenwich Associates 2018 Alternative Data Customer Journey Study

PERSON-HOURS REQUIRED TO EVALUATE ALTERNATIVE DATA SOURCES



Note: Based on 38 respondents.
Source: Greenwich Associates 2018 Alternative Data Customer Journey Study

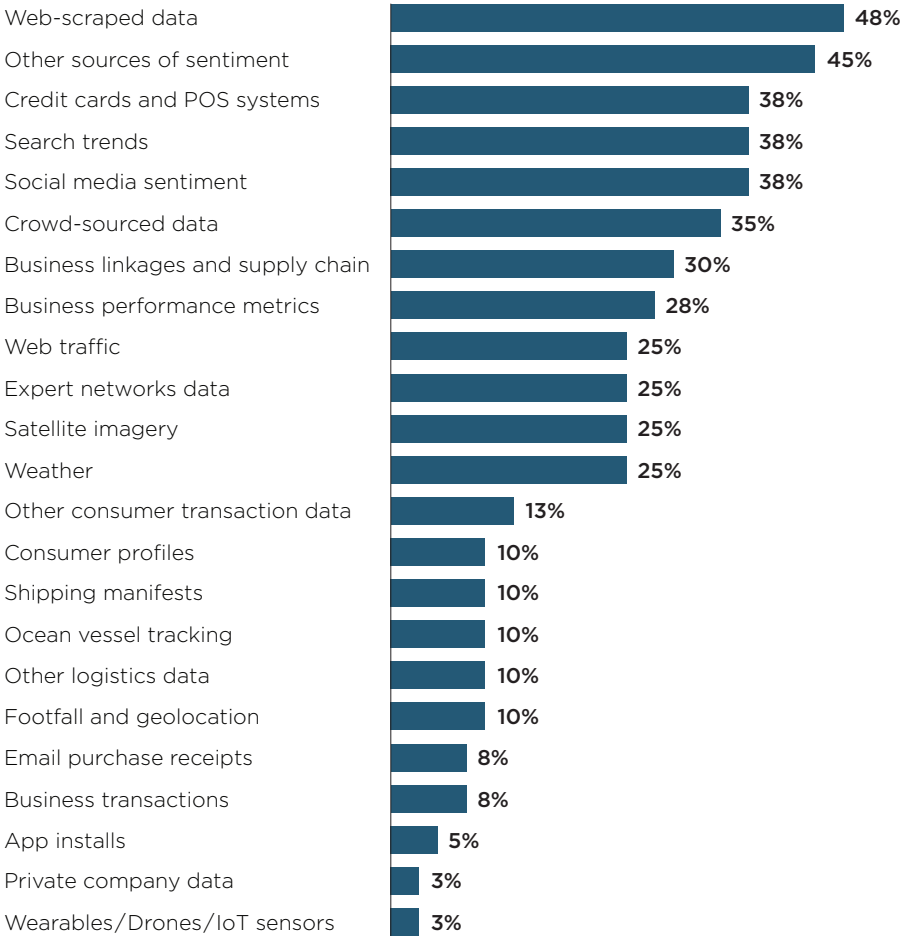
Data is being generated and, increasingly, monetized from sources that would have been unimaginable just a few years ago. Advances in sensors and their reduction in size and expense have made it possible to connect devices to the internet. Lightbulbs can report their individual electricity usage and be aggregated to provide estimates of overall electricity usage. For example, one can now tell how often a warehouse is running night shifts by checking how often the lights are on. Sensors placed on pipelines can give estimates of oil production and distribution figures. These are examples of how companies that are not traditional sources of data can turn data about their operations into an additional revenue stream, potentially turning every company into a data company.

Another source of alternative data is the commercialization of data once generated for another purpose—for example, satellite imagery originally developed for military and scientific applications. Grain companies and commodity traders alike can access live weather data as well as imagery of agricultural land utilization to forecast plantings and prices. Image processing costs have declined, which has improved the resolution so that it is now possible to count cars in shopping mall parking lots and use this data to predict retail sales. The research arm of a major investment bank participating in this study stated that they could predict official economic data with statistically significant accuracy using alternative data-gathering methods.

Mobile phone tracking is a potential source of alternative data but one that also shows the influence of the regulatory environment. The use of cell phone data raises significant privacy concerns, such that its utilization varies significantly by jurisdiction. Geolocation can be used with remarkable precision to monitor customer behavior but requires significant effort to preserve the privacy of the users. For example, an American company anonymizes cell phone users, but classifies them by noting in which census tracts their cell phones spend the night. In

other jurisdictions, this is prohibited altogether. Credit card and point of sale (POS) data are among the most popular forms of alternative data, but here too significant privacy issues arise. Efforts must be made to aggregate or anonymize the data, and the legal usability of this data varies significantly by jurisdiction.

USAGE OF ALTERNATIVE DATA SETS



Note: Based on 40 respondents.
Source: Greenwich Associates 2018 Alternative Data Customer Journey Study

The internet itself is a major source of alternative data. Recent Greenwich Associates data showed that web scraping was the single most commonly used form of alternative data. Search results and social media sentiment are also very important tools. A technology that has recently been perfected is natural language processing (NLP). This has wide application, from document review in law cases to scanning newspapers and Twitter feeds for actionable market intelligence. Another burgeoning use of NLP is the transcription of phone calls. This has a critical regulatory use, as compliance departments at banks are often required to record their employees' interactions with clients to ensure that bank

employees are compliant with the firms' communications policies. These calls can now be transcribed and then searched for key words and phrases that may indicate wrongdoing.

The market for alternative data is new, and this is a pivotal moment for it. The novelty of the technologies and the wild success of a few innovations may have brought many firms into the business. Some firms for whom data monetization is not their main business may decide that the benefits of entering the data business do not justify the costs. For the full potential of alternative data to be realized, the challenges of scaling while protecting privacy rights must be met.

For the full potential of alternative data to be realized, the challenges of scaling while protecting privacy rights must be met.

The early adopters of alternative data have been the most sophisticated quantitative firms, which require little data normalization because they can do that themselves. At the same time, the marginal value of many unique datasets declines with each additional user. So, as firms try to scale, the costs of production rise due to the kind of processing and normalization required that financial data vendors do. At the same time, revenues per client tend to decline as the value to the marginal user declines, and the normalization process itself may reduce the power of signals the data can generate. In this way, the fundamental economics of alternative data work against it scaling.

Regulatory issues are also a factor in the utilization of alternative data. In general, the larger the dataset, the stronger the signal. Accordingly, for many alternative data applications, the individual data points are of little or no use to the consumer. Nonetheless, many of these datasets are composed of what could be personal and protected data, which creates a challenge that regulators and data firms must resolve. The technology at the core of this is globally available, but different jurisdictions apply different regulatory approaches, ranging from limitation to prohibition. How firms respond to the challenge of meeting different regulatory standards will significantly impact how and where the field advances.

What will the future of alternative data look like? The technology that underlies alternative data is globally available, and there are certainly areas in which it can be reconciled with regulatory concerns, so the expansion will likely continue. When thinking about the future of alternative data, it is helpful to think about the maximalist end state and then consider which paths the industry will take in that direction. Given the technology, alternative data, if fully utilized across the globe, may be thought of as a sort of conditional omniscience. Sensors have developed to the point where nearly everything can be monitored and, therefore, everything can be measured. This means that the level of granularity with which human and natural activity can be observed, and thus forecasted, has increased significantly.

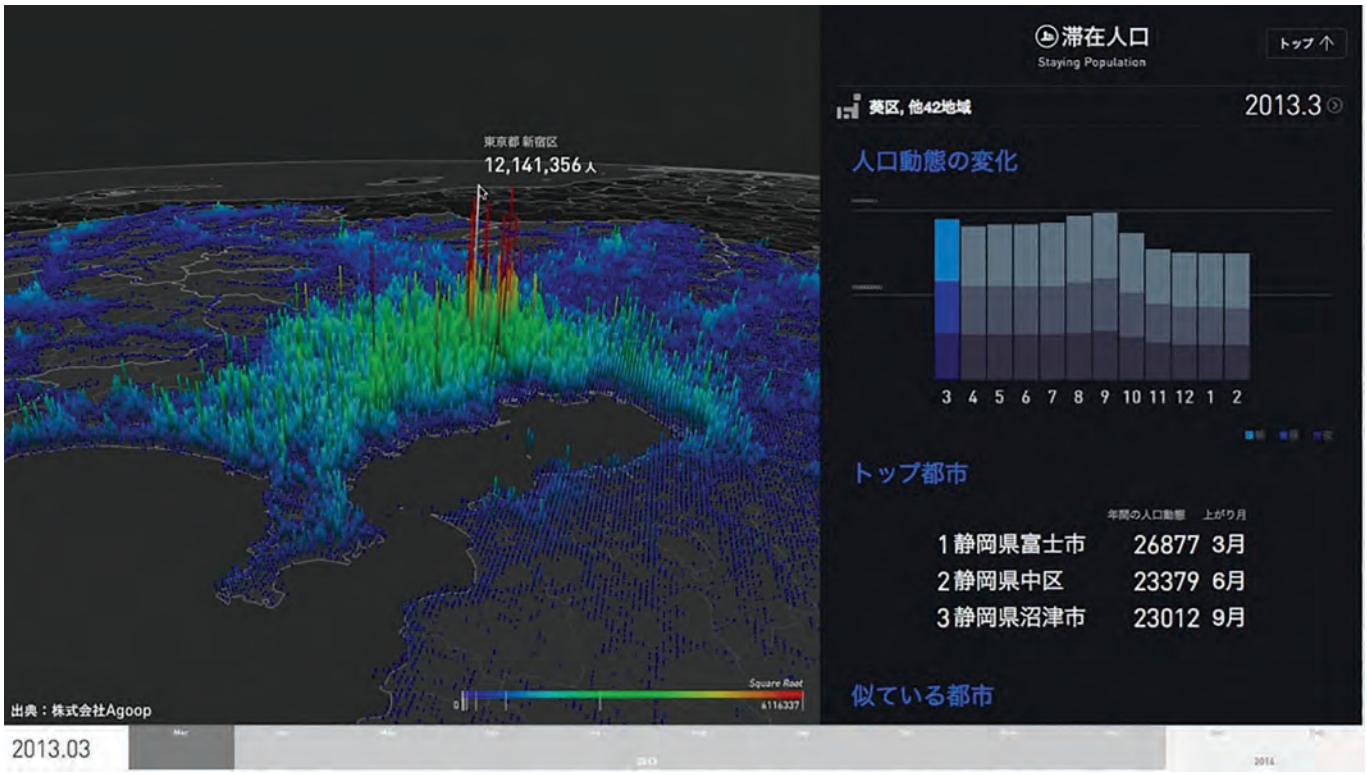
Contrast this with how people measure human activity today. Unemployment reports are generated monthly, company earnings are generated quarterly, much social data is produced annually, the U.S. census takes place every 10 years. All these reports are the cumulative results of activities that can increasingly be monitored as they occur. Thus, firms that are able to invest or procure granular information about these larger reports will have an advantage over those that do not. Over time—and this is the hope of many of the alternative data firms—access to these more granular datasets may become a basic requirement for investors, just as the capacity to interpret public data is today.

While the diffusion of data into markets is highly impactful, perhaps the most important consequence of alternative data will be the fact that it can be acted on. The famous maxim by academic and management consultant Peter Drucker, "If you can't measure it, you can't manage it," may have to be retired in a world in which virtually anything can be managed. Take for example the rationalization of supply chains over the past two decades, which has resulted in "just in time" deliveries and the reduction of working capital tied up in inventories and, consequently, greater economic efficiency the world over. This is precisely the kind of management that can now be done throughout every enterprise. The reduction in the costs as the technology becomes more widely available means that these kinds of management improvements will become available to midsize firms that lack the scale to optimize if the fixed costs were higher.

The policy implications are also significant. Just as corporations manage what they can measure, so too do regulators and governments. Many of the statistics that drive markets are generated by the public sector, and their capacity to manage to them on a more granular level has the potential to confer significant improvements in the quality of life for their citizens. This is true at every level of government. At the national level, a more granular understanding of economic activity can help states direct aid where it is most needed and plan in advance of changes in the economic cycle. There are effects down to the municipal level, where better tracking of vehicles can enable congestion pricing to reduce carbon emissions as well as travel times.

Some governments have set up entire data units. Japan has created the Regional Economy and Society Analyzing System (RESAS), which gathers data from the census as well as from a variety of alternative sources. Regional and municipal governments have been able to use this data for a wide range of uses, from load-balancing medical facilities across a region to developing optimal sightseeing tours. The Singaporean government also has a data unit with an array of applications. One of the more charming and innovative of these is an app for tracking "community cats." Cats are not permitted in public housing units in Singapore, so many communities adopt strays and the Singapore data arm has an app that helps communities track, locate and, of course, photograph their local cats to make sure they are fed and cared for.

REGIONAL ECONOMY AND SOCIETY ANALYZING SYSTEM: POPULATION DENSITY OF TOKYO



Source: Takram 2015[1]

Many alternative data companies would like their products to become the standard, to remove the "alternative" label from their products and make what is "nice to have" today into the baseline. Naturally, regulatory issues and privacy concerns will channel this development, but alternative data could raise the stakes for many businesses, as the level of granularity through which they can view their markets and customers could advance significantly. The "just in time" manufacturing and logistics revolution we witnessed a decade ago could be superseded by a "before time" process, as more granularity in the data permits longer range and more accurate forecasts. As alternative data sources are used more frequently by governments, citizen expectations for what constitutes responsive government may also advance. The possibility that companies and the public sector might be able to discern the needs of customers and citizens in advance of the people themselves raises the bar for competition and for governance significantly.

Still, alternative data usage is relatively new, and its advance is contingent on the proper alignment of technological advance and regulatory prudence. The direction of the industry can be determined by the incentive structure in which its innovators operate. Alternative data will advance the most swiftly along the lines of where the data can confer the most value and where gathering it is the least intrusive.

[1] https://www.takram.com/projects/resas-prototype

Alternative data providers will seek to scale their businesses, which will necessitate the transformation of the data into more easily usable forms. Financial services firms focused on data will be early adopters, but corporations will also increasingly utilize data to optimize their operations. In parallel, governments will increasingly utilize alternative datasets—initially to monitor but ultimately to guide policy decisions. The net effect will be to shorten the time horizon over which decisions can and must be made.

The Cloud and Advanced Analytics

At the core of the data revolution of recent years have been two, mutually supporting technological advances: cloud computing and advanced analytics, popularly known as "machine learning" (ML) and "artificial intelligence" (AI). Both rely on technologies and techniques that have been around for a long time. What is new is the way economies of scale, both in infrastructure cost and in data analysis, have made them much more widely available. As with other aspects of the data world, their advance is channeled both by economic forces and regulatory requirements in the jurisdictions in which they are used.

Cloud computing is simply the utilization of computing power or storage of one computer by another. Originally, the power and storage accessible to any given computer was limited to the hardware of that particular computer. The capacity to connect computers to one another increased their ability to share resources, and the creation of the internet made it possible for computers to share resources with virtually any other computer connected to the internet. This altered the fundamental economics of computing because it created massive economies of scale. The fixed costs of technological infrastructure are very high, but the marginal costs of operating it are relatively low. So the more customers over whom you can spread the fixed costs, the easier it is to finance the higher fixed costs, the more money you can invest.

The capacity to deploy computing power and data storage hardware over the internet and thus spread the fixed costs of infrastructure over massively larger client bases has created a revolution in the accessibility of both.

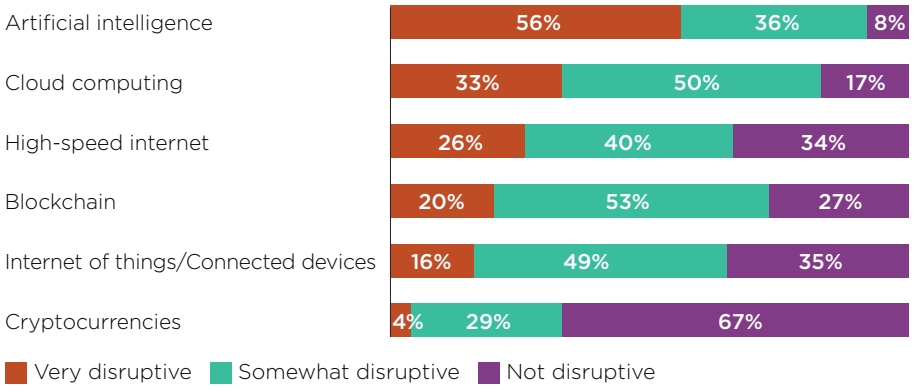
The capacity to deploy computing power and data storage hardware over the internet and thus spread the fixed costs of infrastructure over massively larger client bases has created a revolution in the accessibility of both. Large distributed infrastructure providers specialize in providing these resources at scale to third-party clients. This is also helpful in internet security, a key concern of regulators, because the tools to protect networks also have economies of scale. Large infrastructure providers are better equipped to protect data than atomized firms for which technology is a cost center, rather than the main business.

The declining cost of computing power has led to the widespread use of new tools to process and analyze data. Artificial intelligence and machine learning refer to a set of statistical techniques which are not new per se, but which can be applied in far more sophisticated ways. Fundamentally, AI and ML can be described as automated recursive statistical analysis. Insights that algorithms derive from a dataset can be

reincorporated into the algorithm automatically. That is, the facts derived from the analysis can be used to refine the analysis itself, and this can be done over and over while the model is running—the model uses the data to train and improve itself.

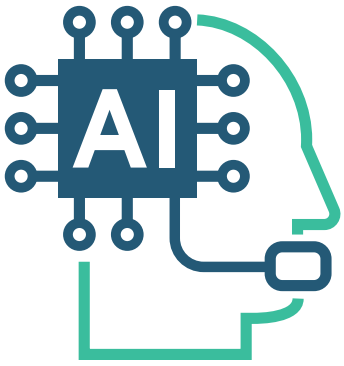
This recursive feature means it is not always necessary for an analyst to define precisely the insight for which he or she is looking. Algorithms can be applied to datasets, with general instructions like "look for similar preferences among groups," and the algorithms can look on their own for patterns and potential forecasts, often leading to insights or relationships that would not have occurred to the analyst. This has significantly increased the capacity for firms to both commercialize datasets whose utility might not have been immediately obvious, as well as to look much more deeply at datasets within which they know there is significant value. These kinds of techniques have been applied to a wide range of problems, from self-driving cars to facial recognition to fraud detection.

DISRUPTION OF TECHNOLOGY



Note: Based on 107 respondents.
Source: Greenwich Associates 2019 Future of Trading Study

The financial services industry has been quick to adopt AI and ML, since large datasets of financial transactions have been available for a relatively long time and so have lent themselves to these techniques. Today, very large numbers of transactions in equity markets are executed automatically by algorithms or "algos." The logic that these algos use to decide how to handle an order, whether to be patient or aggressive, are driven by models that use ML to comb through billions of prior quotes and transactions. Similarly, automated financial advisors are able to recommend portfolio allocations to retail investors using AI by asking a few questions to establish their risk preferences, then comparing them to the universe of similarly situated people and their investment outcomes.



Advances in ML are also facilitated by changes in the way that data is stored. Historically, data has been produced and consumed inside silos. This is logical—a business unit would produce data for its own use, store it and analyze it when necessary. Other business units might not even be aware of its existence. As AI and ML techniques began to produce results, firms have been looking for ways to increase the scope of the data to

which they can apply them. The cloud not only increased the availability of computing power, it also increased the accessibility of data storage capacity. Once firms started moving their data to the cloud, the concept of a "data lake" was born.

A data lake is a remote facility that enables a company to store and access its data remotely, regardless of the structure or format. These datasets can be made accessible to all the other parts of the company, breaking down the data silos that had existed when the data was stored onsite or even in a highly structured data warehouse. For many firms, this pooling of multiple datasets combined with the use of ML techniques has been revolutionary. Storing data in the cloud evokes something ephemeral and distant, but this is the wrong way to think about it. On the internet, the cloud is a fixed location and immediately available. As a study respondent said, "That's the thing about a data lake, you can't move a lake. You have to go to it." This is a key insight: The data is now stationary, and the analytics can be brought to it.

"That's the thing about a data lake, you can't move a lake. You have to go to it."

Moving large amounts of data is expensive, time-consuming, error prone, and, depending on the data and the jurisdiction, may create significant privacy and legal issues. The technical issues are best understood with an example: Refinitiv has two petabytes of tick data going back to 1986. For reference, two petabytes of mp3 files would take a person 4,000 years to listen to. To download only the FX tick data from their database would take three months. This is no longer necessary, however. The data is in a fixed location, and firms can bring their ML tools to it. Moreover, multiple processes can be run on the same dataset simultaneously. You can see these effects in entertainment: People no longer download music and films, they stream them. The business model of centrally storing massive amounts of data and running multiple client processes on it simultaneously has conquered the media industry. This change is just getting started, and the consequences will be significant.

An Outlook

To understand where this is leading, it is important to be more specific about what the word "insights" means when talking about how data is used. The results of data analysis take a number of forms. Analysts look for patterns and relationships among elements and subsets of the data. Significant patterns and relationships often reveal themselves through the analysis of large datasets that would be invisible to someone looking at individuals or smaller groups.

Taking this a step further, analysts look to establish causal relationships. They try to determine whether one set of circumstances can be linked through time to another set of outcomes. Then, when new data comes out, they are able to predict what is likely to come next. A useful feature of this kind of data analysis is that there are tests to measure and terms to describe the predictive power of a particular analysis. Certainty is

extremely hard to come by, but being able to make an educated guess and have a quantitative measure of how likely your guess is to be true, is as close as humanity can get to observing the invisible or seeing into the future.

The advances in ML and their application to massive datasets at low cost by leveraging cloud computing represent a fundamental change to our epistemology. Epistemology is the investigation of what distinguishes justified belief from opinion. Using these new technologies, virtually any question that can be asked about the relationship of any two things for which a large enough dataset exists, can be determined—perhaps not with certainty, but with a level of confidence that can be measured. This knowledge can then be used to inform decisions that will be significantly better than those made through guesswork or intuition. This is especially important for causal relationships, because it means a much wider variety of things can be predicted, reducing the overall level of uncertainty. Given this technology, the moment one can frame a question in these statistical terms, the only technical barrier to answering it is whether sufficient data on it can be collected. This is truly a unique moment in intellectual history, comparable to the triumph of experimental science over received wisdom in the 17th and 18th centuries.

The advances in ML and their application to massive datasets at low cost by leveraging cloud computing represent a fundamental change to our epistemology.

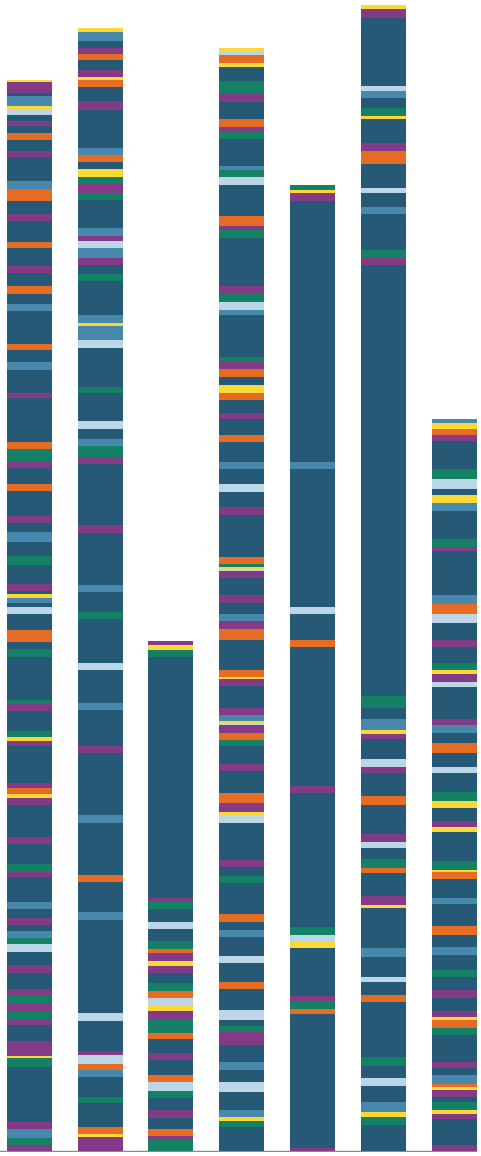
This expansion of what can be known statistically is being put into practical use throughout finance. Decisions about creditworthiness used to be based largely on personal relationships and then on extensive documentation. Now the necessary information can be accurately inferred from data that might not seem to have any direct relationship to the decision. Foreign exchange liquidity providers are able to determine within a handful of trades the nature of the client, the risks involved in dealing with it, and minutely shift their pricing to account for it. Regulatory authorities have been able to use these techniques to identify insider trading simply from anomalous patterns in transaction data, rather than through direct surveillance of a suspect. Central banks are using these techniques to look over the horizon and see risks before they arise. Sophisticated financial services firms continue to be at the forefront of utilizing these tools to improve their pricing and risk management techniques. The wider implications of this are profound.

As a thought exercise, imagine if an academic institution were granted simultaneously all the data and all the computing power on Earth, as well as the capacity to deploy alternative data-gathering methods for any datasets which it did not possess. Imagine this university was then given the task of improving human life. Such an organization would be able to find patterns and links between phenomena that would have been invisible to any analysis but this. It might establish robust causal relationships between things that would have been impossible to determine before this technology existed. From a statistical perspective, there would be virtually no questions it could not answer, even if only

in the negative. Of course, it would not be perfect—all of its predictions would have to be qualified with a confidence measure determined by the quality of the data—but human capacity to make informed judgements would be significantly increased.

The range of potential applications of these techniques is as broad as the universe of data that might be collected. How would healthcare outcomes improve if all the results of every procedure to address every illness were known? How could crime be reduced if we knew in advance both who was most at risk of victimization as well as what interventions were most effective at reducing crime or recidivism? These techniques could also be applied to the natural world. Wind patterns could be analyzed, better predicted and wind farms better sited to improve clean power generation. The lifecycles of undersea life could be better modeled to improve fishing sustainability. As the costs of acquiring data, storing it and analyzing it drop, the power to answer these kinds of questions, with levels of certainty that can be measured, are increasing accordingly.

It is important to remember, however, that this is a human and fragmented world, so these resources will continue to be used in pockets. Therefore, to improve their usefulness, people will have to cooperate. The thought exercise is helpful to understand the immensity of the latent power of the technology, and this is the direction in which the future of data is going. Of course, there are constraints that are more important than the technological. The need to respect human rights might require laws to limit the gathering and utilization of some datasets. Additionally, data itself has value, and the owners of data expect to be compensated for it. As a result, the trajectory of the data revolution enabled by the advance of machine learning and improved data storage will, like other elements of the data world, move fastest where the conflicts with regulations will be the least and where the financial rewards will be the greatest.



As mentioned, the early adopters have been the most sophisticated firms in finance, both banks and investment managers. That said, the nature of the cloud is such that it enables small and midsize firms to access technologies and methods that would have been out of reach historically. As a result, many more firms will be able to optimize their businesses and discern patterns within their data than would have been possible before. Governments and academia, which also possess access to enormous datasets, will also utilize these techniques as they are pioneered. Every entity in the world is moving toward a conditional and local omniscience circumscribed by the datasets to which it has access, the economic value it can generate from them and by the regulatory regime within which it operates.

Case Study I: How Can Data Fight Crime?

The adverse impact and harmful consequences of financial crime are massive. The United Nations Office on Drugs and Crime estimated in 2009 that criminal proceeds amounted to 3.6% of global GDP and that approximately $2 trillion in illicit proceeds is laundered through the financial system annually. These figures would place the global criminal enterprise as the eighth largest economy in the world, in place of Italy. Moreover, the United Nations estimates that a mere 1% of criminal funds that flow through the financial systems globally each year are detected and subject to law enforcement action.

Financial crime is a hidden, illicit economy. It is structured into highly complex, opaque networks. This complexity and opacity, as well as its frequent attempts to blend in with the legitimate economy, makes it extremely difficult to identify and combat.

This is a huge challenge for the authorities as well as for the financial services industry. Financial services firms are legally and regulatorily obligated to ensure that they put measures in place to combat financial crime. Severe penalties have been levied against firms that have failed in their duties in this regard. In addition to the large sums of illicit money generated by financial crime (including fraud, human trafficking, narcotics crime, and terrorism financing), the significant negative social impact must be considered. For example, despite the fact that every country on the face of the Earth had banned slavery by the end of the 20th century, it has resurfaced in the 21st. The International Labor Organization has estimated that 40 million people are enslaved today.

Financial criminals need to legitimize their criminal proceeds and financial services need to be able to identify these illicit funds and their sources in order to comply with legislation and regulation and play their part in preventing financial crime.

In the context of financial crime, it is widely acknowledged that money laundering—the process of making illegally gained proceeds appear legitimate—generally comprises three steps:

1. **Placement**—entry of the illicit funds into the legitimate financial system
2. **Layering**—disguising of the illicit funds through a series of obfuscating transactions
3. **Integration**—making "clean" funds available for use from a seemingly legitimate source

Combating money laundering and financial crime also includes three steps:

1. **Know your customer (KYC)**—It may seem self-explanatory, but this process can be quite complicated. Many clients of financial services firms are not individuals who can be easily identified but are often corporations or partnerships whose ultimate beneficial owners might be difficult to discern. Additionally, what constitutes an anti-money laundering (AML) risk to institutions is quite broad and diverse. For example, with regard to financial crime involving state funds, the recommendations of international agencies such as the Financial Action Task Force (FATF) include the need to ensure a necessary risk assessment is done before conducting business with state-owned entities and "politically exposed persons" (PEPs). Conducting the required level of KYC due diligence requires access to trusted, quality data sources to help determine the risk, if any, associated with a particular customer.
2. **Transaction monitoring**—Once a financial institution knows its customer, it must still monitor their payment and transfer activities in line with their legal and regulatory obligations. Most customers of financial services organizations have established patterns of activity, so a departure from these regular patterns can indicate that there is a potential risk of money laundering associated with these transactions.
3. **Payment screening**—Finally, an apparently legitimate customer may be sending funds to a third party, which may represent a money laundering risk. It is therefore important not only to know your customer, but often to know your customer's counterparties as well. This is done to prevent both layering and integration.

The successful prevention of financial crime requires financial services organizations to answer a series of complicated questions, including:

- How to determine whether a given customer is on a sanctions list?
- How to know whether they have committed a financial crime in a jurisdiction other than the one in which the firm operates?
- Who has taken political office and now constitutes a politically exposed person?
- Are there degrees of political exposure?
- Which customers merit more careful monitoring than others?

In order to help financial services firms manage risk and answer these questions, access to trusted data from a wide range of sources is needed. However, much of the data—including official government, law enforcement and regulatory, as well as media sources—is unstructured and requires collating, normalizing and structuring to make it usable and

to ensure accuracy and relevance. This is the work that financial data firms like Refinitiv perform with public domain data and include in the World-Check Risk Intelligence Database.

World-Check's research analysts, adhering to strict inclusion criteria, distill the information into structured records for use by customers as part of the due diligence checks they conduct. These analysts speak more than 60 languages among them and are responsible for reviewing local public domain data to ensure relevant information is accurately captured in World-Check.

Specialist teams work on specific areas of financial crime risk, including sanctions and regulatory updates, organized crime, environmental and human rights crimes, and PEP data. The value of human expertise to recognize, and include in World-Check, the often complex connections between certain individuals and companies associated with financial crime networks is critical to providing the holistic picture that allows customers to assess potential risks that may lie within these networks.

Other datasets, such as vessel data and ultimate beneficial ownership data, have been integrated within specific World-Check services to enable clients to identify and verify client information and then screen against the World-Check Risk Intelligence data to ascertain if there is any associated risk in one smooth workflow.

Automation and AI are used to supplement the efforts of human researchers in their day-to-day activities. Moreover, tools such as an AI-powered media screening feature, which uses new technologies including NLP and ML combined with specific AML and financial crime taxonomy and intelligent tagging, can be used to deduplicate and group relevant media articles to boost operational efficiencies within an organization's KYC due diligence processes.

The fundamental value of World-Check is the ability to structure unstructured data. It does this by combining advanced data techniques with human intelligence to scrutinize data from thousands of sources, connect intricate networks and reveal hidden associations. This provides financial services firms with a holistic picture of their potential risks and assists them in their obligations to comply with financial crime legislation.

That golden source is large and getting larger. The World-Check database contains millions of records. The database is used by financial institutions, corporations, professional services firms, governments, law enforcement agencies, regulators, and others to perform due diligence and other screening activities in accordance with their legal or regulatory obligations and risk management procedures. The database continues to grow and is regularly updated.

Harnessing the power of data can make a significant impact in tackling financial crime, terrorism and modern day slavery.

World-Check combines advanced data techniques with human intelligence to scrutinize data from thousands of sources, connect intricate networks and reveal hidden associations.

Case Study II: Can Better ESG Data Reorient the Financial System to Advance Sustainable Development?

In recent years it has become increasingly clear that humanity faces challenges on a scale that cannot be addressed by anything short of global cooperation. Climate change is the most salient of these, and the global community has come together in a variety of forums seeking to address it. The one thing that all potential methods for addressing global challenges share is extraordinarily large investment requirements. The United Nations has estimated that investments of $5–$7 trillion will need to be made to achieve the Sustainable Development Goals (SDGs). Meeting these objectives will require the full mobilization of the global financial system. Two phenomena, in turn, are key to this successful mobilization.

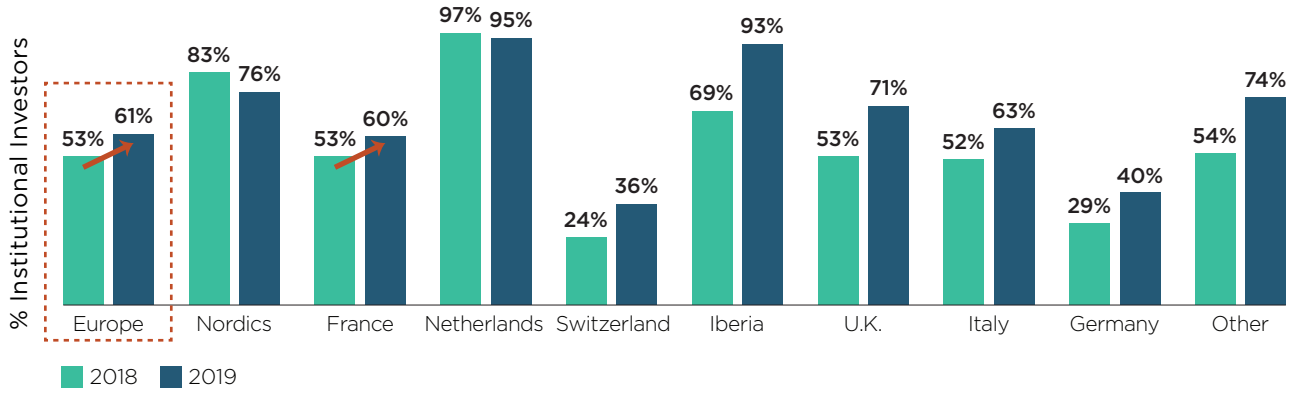
The first is the shift among investors toward socially responsible investing (SRI). The movement of investors who want to channel their savings toward investments that provide social good in addition to financial returns is well under way. Of course, effective SRI requires a means of measuring social impact. This has led to the creation of distinct datasets focused on firms' environmental, social and governance (ESG) strategies and their implementation. ESG data is security- and firm-level fundamental data. Although related, SRI and ESG are distinct concepts. SRI relies on and provides an impetus to the ESG data ecosystem, but ESG data would be valuable with or without SRI. This is because it measures the impact of corporate activity on a wider universe of stakeholders and can help companies optimize their operations to achieve wider social goals.

The speed with which socially responsible investing has swept the financial system in the past few years is nothing short of incredible. A study participant said it best when he described the financial system as "permeated" by the shift. At Greenwich Associates, we observe this increased level of interest in conversations we have throughout the financial system. Investment managers see it as a way to respond to investors, the sell side sees it as important to the managers and seeks to create products and services that support it. Exchanges and index providers are focused on providing products to the sell side. There is no corner of finance that has been untouched by SRI.

The evidence is not simply anecdotal—it is evident in our data. In a study of nearly 800 asset allocators across the United States and the EU, 60%

said that the socially responsible investing approach was an important consideration in their selection of asset managers, up from 53% just a year ago. A common misperception is that millennials are driving the ESG trend. While it's true that interest in ESG is highest among this group, at 77%, majorities of both Generation X (64%) and Baby Boomers (61%) were also interested in SRI. The interest in SRI is neither limited to a particular demographic nor to a particular geography. It is driven by an increased level of demand among investors of every kind that choose to channel their funds and think more deeply about how they do so.

SRI CONSIDERATION FOR MANAGER DECISIONS



Note: Based on 734 respondents in 2018 and 736 in 2019.
Source: Greenwich Associates 2018 and 2019 Continental European and U.K. Institutional Investors Studies

To date, much of the focus of SRI has been on the passive side of the equation. The first steps forward in SRI were using what ESG data was available to set up screens and develop indexes of companies that met certain ESG criteria. ESG data at the time was relatively new and was able to return a binary "include" or "exclude" decision, which lent itself toward creating indexes. This was only the beginning.

As ESG data quality progresses, however, it will increasingly be used to generate alpha. The social component of ESG is often interpreted as how well firms promote diversity and inclusion within their ranks and among their executives. Academic research has been finding improvements in the results generated by firms that have relatively diverse management teams. Effective inclusion policies may prevent group think and promote intellectual diversity. The G in ESG stands for governance, and a number of high-profile issues at technology companies, as well as the history of large regulatory fines, make a strong case for the link between governance and returns.

There are challenges standing between investor interest in socially responsible investing and its successful implementation. While there is a great deal of interest in implementing SRI strategies by investors, what constitutes SRI investing is very much in the eye of the beholder. There are no uniform standards for what constitutes a desirable SRI investment nor market practice standards on how to rank or even to judge them.

The absence of official standards is not fatal to the project, however, as investors have been deciding for themselves what constitutes good corporate strategy and performance, though in this they have been helped by the required disclosure of financial data. The fundamental challenge to improving SRI is an ESG data challenge.

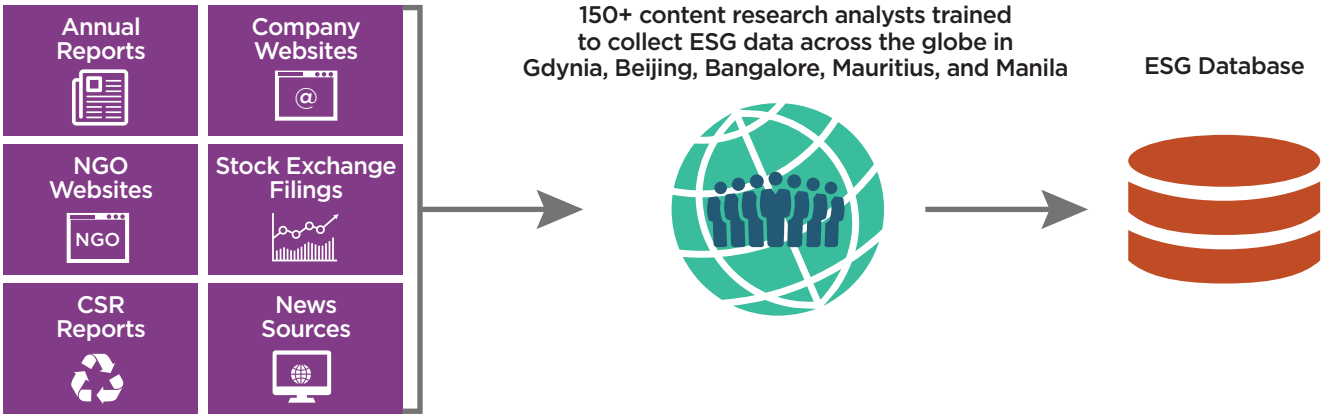
Firms are working today to identify and source datasets that can be used to make estimations of ESG data and aggregating them into a form that can be used to guide financial decision-making. Some data is available from required corporate disclosures or board-certified CSR reports, although at the moment, there are few officially required ESG data disclosures. As a result, investors are compelled to rely on other, non-standard data sources. Many of these datasets are unstructured and require sophisticated tools to properly utilize and normalize them.

These data aggregation tasks play to the core strengths of financial data companies, and these firms, in competition with some specialized data companies, have worked to productize it. For example, Refinitiv generates ESG scores for over 7,000 firms covering 70% of global market capitalization. They gather data from structured datasets like corporate filings and add data collected from other sources such as web scraping, news reports and reports by NGOs. Once they have gathered the data, they clean and distill it to illuminate corporate performance on hundreds of individual measures. From these they generate scores measuring performance on subcategories of the ESG universe as well as an "ESG controversy" score, a measure of media coverage of controversies in which the company is concerned, which may have an impact on ESG-focused investors. These scores are weighted and combined into an overall ESG score.

The future of ESG data will be driven by the need to overcome the shortcomings of the ESG data universe as it stands today—mainly issues in the raw materials that data firms have to work with. Much of today's ESG innovation is focused on how to gather and utilize data. There remains a trade-off between accuracy and completeness. The more factors an investor wishes to consider, the less accurate their measures are likely to be. At the moment, ESG data is robust enough to establish indexes, and it is beginning to be useful in generating alpha by observing marginal differences between firms. What is necessary to secure the potential for SRI and ESG to mobilize the financial system in the service of wider aims is the ability to measure outcomes. This is the future of ESG data: the improvement of the raw materials of ESG data to pave the way for better analytics and more robust linkage between investment and outcomes.

This is the future of ESG data: the improvement of the raw materials of ESG data to pave the way for better analytics and more robust linkage between investment and outcomes.

ESG DATA REFINEMENT AND COLLECTION



Data quality is a key part of the collection process; that is why we use a combination of both algorithmic and human processes to make sure we achieve as close to 100% data quality as possible. Below is an overview of the various methods we use to achieve this goal.

Data Entry/Pre-Product

- Around 400 built-in error check logics in the collection tool for various data points
- Error checks can be tailor-made for specific requirements

Post-Production

Around 300 automated quality check screeners run on ESG collection tool:

- Relating to interrelated data points
- Negative screening
- Inconsistency or missing in quants and qualitative
- Scaling
- Variance within a year
- Raw data and comments section
- Sector-based checks (TRBC codes)
- Validating completeness of the prior year

Independent Audits

Sample audits on a delay basis:

- Detailed audits
- Critical data point checks
- Product audits

Weekly reporting and root cause analysis:

- Feedback sessions with the production teams

Management Reviews

- Monthly quality—deep dives
- Heatmap analysis with top areas for concerns
- Measures to address problematic topics and data points
- New system validation checks and screeners are constantly created based on new learning, insights and feedback to continuously improve the data quality

Source: Courtesy of Refinitiv

To see the future of ESG data, it is helpful to look into the past of fundamental financial data. It could be said that ESG data today is where traditional corporate fundamental data was in the United States before the passage of the Securities Exchange Act of 1934. Prior to the Act, there were some disclosure requirements set by exchanges, but they were pro forma and hobbled by a lack of accounting standards. Some corporations voluntarily disclosed financial data, but they had no obligation to do so and could be selective and irregular about it. As a result, the true nature of the financial conditions of many corporations were known only to a handful of insiders with privileged access to the

firms' accounts. This created significant opportunities for executives to enrich themselves at their investors' expense, which undermined confidence in the markets as a whole.

During the Great Depression that followed the crash, the political will was created to change this permanently and make the markets transparent and more open to all. The '34 Act established uniform standards for all U.S.-listed companies to disclose data relating to their financial health on a quarterly and annual basis. These disclosure rules were driven by the concept of materiality. They defined what constituted material and non-public information and strictly regulated its use in order to restore investor confidence and combat the Great Depression. Today, ESG data has the status of financial data in 1925. It exists, it is material and it is non-public.

To get a sense of just how material it is, conduct the following thought exercise. Imagine for a moment that all of the effects of the next 10 years of climate change, both the natural and the regulatory responses to its intensifying effects, were telescoped into a single year. This would obviously be an incredible calamity but leaving aside the horror of this, imagine two different worlds: the current one and another in which ESG data is as accessible to the public as financial data is today. In the current world, investors would struggle to understand the effects of the disaster and might panic and abandon the markets altogether, hindering efforts of governments to combat the crisis.

In the second world, investors would have unfettered access to all the potential ESG data for all the companies in the world: their physical exposure to the effects of climate change, the effects on their supply chains of potential disruptions, and their carbon intensity and, thus, their sensitivity to increases in carbon taxation. In this world, investors would be able to gauge quickly which firms were most at risk and which firms had climate policies and carbon intensities that would be more resilient to the crisis. Additionally, they could more easily assess which firms had potential solutions to the calamity. They would be able to take all these issues into account when reallocating their funds. Thus, the private sector could help accelerate the course the regulatory authorities would be plotting to combat the crisis, rather than slowing it by withdrawing in fear of the unknown.

Given the difference between these two worlds, the key question for the future of ESG data is, why wait? Though the effects of climate change may take years to manifest themselves rather than months, ESG data is no less material. Indeed, it is more so. What is at stake is whether investors are actually able to measure if their SRI decisions are having the desired effect, as well as the capacity for corporations to optimize their non-financial performance in line with their ESG, as they currently seek to optimize their financial performance today. One could argue that the stakes with regard to ESG data are higher than they were in 1934, a lot higher.

This lack of disclosure requirements is not the only challenge to the mobilization of the financial system in the service of wider goals. Determining just what should be disclosed is also a challenge. In the same way that the objectives of SRI investing are to some extent in the eye of the beholder, ESG data is also somewhat subjective. This makes for many open and extremely complex questions. What is the standard of materiality that should be applied? What should the metrics for carbon intensity be? How should carbon intensity in the supply chain be measured? What constitutes a successful implementation of a diversity and inclusion policy? To what threshold should the potential impacts of climate change be factored into risk statements? Can they be quantified? If so, can they be quantified in a uniform way that permits comparison among firms? Will they be consistent through time so that firms can measure their own progress?

This complexity need not be a hurdle to bringing the future of ESG data closer to the present, however. Just as disparate groups have come together under the auspices of the UN to establish the Sustainable Development Goals, it should be possible for interested parties to agree among themselves what should constitute ESG data. Analogous tasks were performed by the authors of the '34 Act, as well as by the Kyoto negotiators. The questions are complex, and some answers may have to be reached through compromise rather than consensus, but what is needed are answers.

This process, the collective determination of what constitutes the most vital ESG data and the implementation of disclosure requirements to ensure universal access to that data are the future of ESG data. They will improve and enrich the processes that data companies have already begun, and they will enable the development of further analytics—both to enable investors to determine how best to direct their capital, and also for firms to determine how they might best alter their operations to lower their costs of capital by attracting investors who share their values. The importance of arriving at clear definitions of what constitutes ESG data and creating a disclosure regime for it cannot be overstated. Repeating the Peter Drucker mantra, "You can only manage what you can measure." Thus, to mobilize the financial system to address climate issues, we must be able to create a standard with which investors can punish failure and reward success.

The collective determination of what constitutes the most vital ESG data and the implementation of disclosure requirements to ensure universal access to that data are the future of ESG data.

Regulatory Considerations for the Future of Data

The future of data is not merely a question of whether economies of scale can be captured or how rapidly technology will advance, it is also a question of how the rules that govern the use of data change and adapt to the new technology and economics of data. The data revolution

has been enabled by the relatively open societies in which it originated, which have long traditions of free expression and the free transmission of knowledge. The most important enabler, the internet, was originally created as a collaboration tool within academia, a sector which has been sharing knowledge for centuries. It is equally important to remember that it was originally financed by a grant from the U.S. Department of Defense. Governments, through their regulatory authorities, will play an extremely important role in the future of data, both because they support the technology and because they determine the rule set in which both the technology and the economics interact.

Financial data is an excellent example of the constructive role that regulatory authorities can play. Much of the fundamental and market data that constitutes the financial data universe is the result of regulatory requirements. Nations have an interest in the efficient allocation of resources within their economies. This requires that investors have the data they need to make decisions about how best to invest their funds. Trust is also a key element of an efficient financial system, so financial regulators have an interest in supporting fairness and equal access to financial data. To this end, regulators establish disclosure requirements both for companies that apply to issuers of securities and to transactions in them. Having established these requirements, regulators need to work continually to increase the quality and reliability of material data. This concept of materiality is important in most global financial regulatory regimes. A company issuing securities on a capital market is required to disclose all material information—data—to the public so that investment decisions can be made. But what is material? What data in this scenario is decision-ready?

At the base of it, regulators have to balance the need for completeness with the need for accuracy, and the costs of producing the data with the benefits of consuming it. ESG data is a good illustration of this. Some elements of ESG data, particularly in the governance realm are required disclosures. Investors must understand how decisions are made and who is responsible for them if they are going to be able to hold them accountable. Other aspects are not so clear. Environmental data is much harder to get, the structure of a company's operations and their carbon intensity can be estimated, but these estimates can vary widely and be difficult to validate. Even if required, they might be expensive for a company to calculate on its own. Different disclosure requirements would have different costs and confer different levels of transparency. This is also a challenge in financial data. Regulators have to make decisions, sometimes arbitrary ones, about what level of granularity balances the needs of investors with the costs of reporting. Nonetheless, having the regulators as a neutral third party, with the power to set and enforce standards, is an effective way to resolve this challenge.

In addition to determining what must be disclosed, regulatory authorities also have a role to play in determining what remains private. The advance of alternative data-gathering techniques and the ease with

which usage and communications over the internet can be tracked have reduced the costs of gathering data. Advances in data storage and analytics have increased the value of access to this data by creating more commercial uses for it. The increased gathering and use of potentially personal data has raised privacy concerns in many societies. Regulators have a responsibility to protect the rights and interests of their citizens, and many are working to regulate the way in which the data industry operates. These regulations will have a profound effect on the future of data.

Societies understand privacy differently, and these understandings affect how they regulate it. In the EU, privacy is considered a fundamental human right enshrined in Articles 7 and 8 of the Charter of the Fundamental Rights of the European Union. This right inheres in the citizens of the EU and they cannot be alienated from it. By contrast, the Bill of Rights of the United States simply prohibits the state from executing unreasonable search and seizure. The U.S. Supreme Court has established a right to privacy, but the effect of this is to bind the state, not the citizen. In Asia, a diversity of views prevails, and the needs of national security are also often at the forefront with regard to considerations of privacy.

These fundamentally different conceptualizations lead to fundamentally different means of regulating privacy and the commercialization of data. In the U.S., the data markets have been relatively lightly regulated and the commoditization and commercialization of data has proceeded apace. By contrast, the EU has passed the General Data Protection Regulation (GDPR), a horizontal data privacy law impacting all sectors. Among the many things the GDPR does is define what constitutes personal data. As with disclosure, definitions that are common across the ecosystem are helpful because they set standards and provide clear limits that enable firms to understand precisely what their obligations are.

In the case of the GDPR, firms have a number of obligations with regard to how they handle and utilize data and EU nationals retain certain rights to their data. This includes the "right to be forgotten," that is, citizens have a conditional right to have their data erased. This and other aspects of the GDPR therefore impose requirements on businesses using the data of EU nationals. Other societies have taken different approaches. Singapore has also enacted its own Personal Data Protection Act (PDPA), and it takes a more principles-based approach. It, too, defines personal data and focuses on individual consent, the disclosure of the purposes for which the data is being gathered, and that the purpose meets a reasonableness standard. When firms are making investment decisions, regulatory certainty and the costs of compliance are factors they take into consideration. In this way, data protection laws can channel the development of data businesses.

Data is a global business and the GDPR had implications far beyond the EU. For virtually any datacentric task, the larger the dataset, the

Fundamentally different conceptualizations lead to fundamentally different means of regulating privacy and the commercialization of data.

more effective the analysis. Therefore, any firm utilizing the data of EU nationals, regardless of its own jurisdiction, has to comply with the terms of the GDPR or face fines. The privacy elements of the GDPR and other data protection regimes raise the costs of compliance and thus deter or reduce certain kinds of investment—but they do not significantly threaten the future of data. There is another form of data regulation that does: data localization.

Data localization laws set geographic boundaries for the utilization or storage of data. They take many forms—some require local copies of data be maintained, others mandate that transmission, processing or storage occur only within the local jurisdiction. Nations enact them for a variety of reasons, but they often struggle to achieve their intentions and almost always result in unintended and often very negative consequences for the country and societies in which they are enacted. The data ecosystem has its own geography, which is driven by the technology and economics that underlie it.

Take the privacy issue: A data localization law that would require copies of data be retained in a particular jurisdiction forces data companies to duplicate datasets. This is counter to the principle of data minimization, whereby organizations seek to limit the personal data they collect. Additionally, the existence of multiple copies of the same datasets in multiple places actually makes it harder to manage and protect, producing the opposite of the desired outcome. Firms that operate in multiple jurisdictions need to be able to track their operations using data from multiple sources. Attempts to confine data locally obstruct this and create incentives to move away from countries with localization laws. Advanced data technologies like machine learning require large datasets to be effective. Data localization laws that require data segregation by national boundaries weaken the utility of these tools. Cloud storage and analytics are also impaired by data localization laws and, in some cases, make these activities economically unviable. Studies are increasingly showing that the direct costs and unintended consequences of data localization greatly outweigh the benefits, many of which never materialize.

The data ecosystem relies on the free movement of data, and the nature of analytical tools confers advantages of scale that localization laws can diminish. The data economy is an economy of openness and interdependence. Data "autarky" or self-sufficiency is a contradiction in terms. Nevertheless, these laws are spreading and are either in force or under contemplation in 84 jurisdictions around the world. The EU's GDPR and the Singaporean PDPA have shown that there are many tools at hand for those concerned about securing the rights of their citizens. Data localization laws are not an effective way to secure the rights of citizens, impose severe costs for uncertain rewards and should be reconsidered.

The future of data will be determined largely by the advance of technology and the economics which that advance enables, but these are not the only factors. Its advance will also be channeled by regulatory

authorities seeking to secure both the welfare and the rights of their citizens. Regulators have a role to play in safeguarding the collection and use of data where it would conflict with the fundamental rights of their citizens. The data revolution will advance, and the utility of the techniques is so great and the cost of accessing them is declining so rapidly that firms will be continually finding ways to put data to work for them. Regulators can help ensure this is done in an ethical and efficient way by safely enabling rather than creating inefficiencies and barriers, thus having a profound impact on how the future of data unfolds.

Conclusion

Because the future of data will not look like the past, data about the past cannot be used to inform it. The spectacular increase in the collection of traditional data as well as new alternative methods of gathering data has made it possible to describe quantitatively more of the world than ever before. Contemporary alternative datasets will increasingly become the norm, even as new alternative datasets are developed and metadata becomes normalized as well. Combined with storage technologies that permit enormous scale and encourage the elimination of silos, the universe of usable data is expanding geometrically.

If you can reduce any question to a data question for which there is accessible data, you can have an answer.

The possession of data and the effective utilization of data are two distinct things, and as the tools for gathering, storing and accessing data have advanced, so too have the tools for analyzing it. Artificial intelligence and machine learning have enabled these datasets to be understood and utilized more completely than ever before. The process of gleaning insight from these datasets only partially requires human intervention, as the algorithms now train themselves.

These advances, and the ease of their application to large datasets, have enabled people to look at and take actions on the phenomena of the world with greater depth and precision than ever before. The very technologies that enabled this scale have reduced costs to the point where these kinds of tools and analyses are widely available. Think of the benefits of alternative data gathering, natural language processing, geolocation, and predictive analytics that occur every time a cab driver asks his or her phone for directions for the fastest route to your home. The spread of access to data and its means of collection have also awakened the authorities to new challenges. They need to ensure that regulations are written so that the benefits of innovation in data can be collected while the rights and privacy of citizens are protected and that new data that is deemed material, such as ESG data, is disclosed. Much will turn on how regulators decide to achieve this.

The data ecosystem is at an inflection point: The parallel advance of complementary technologies and their accessibility to a wide range of actors is making the knowledge and understanding of data a requirement. Greenwich Associates data shows that among financial services firms, facility with data is now the skill most in demand. The potential for progress is incredible. Better utilization of data can help prevent financial crime, allocate society's resources more efficiently, and it can help the financial system achieve the essential task of financing the Sustainable Development Goals. What the advance of the data ecosystem has done is fundamentally expand the knowable. If you can reduce any question, whether about human nature or nature itself, to a data question for which there is accessible data, you can have an answer.

Better utilization of data can help prevent financial crime, allocate society's resources more efficiently, and it can help the financial system achieve the essential task of financing the Sustainable Development Goals.

Appendix: What is Financial Data?

Financial data refers to a wide spectrum of data and may have different meanings dependent on the context. For the purposes of this paper, it may be helpful to consider three broad categories of financial data: fundamental data, market data and reference data.

Fundamental data describes the fundamental state of affairs for an actor in the financial system. Financial actors can be companies, governments or even economies as a whole. Each category of actor has its own forms of descriptive fundamental data. For corporations it might be their revenues or earnings, the number of employees they have or the identities of their officers, or measures of their performance along their environmental, social and governance (ESG) metrics. For governments, fundamental data includes things like tax revenues, expenditures, outstanding debts, term lengths, or constitutional commitments. Many kinds of fundamental data can be used to describe economies as a whole—unemployment or power usage, for example. Aggregated fundamental data of corporations and governments is fundamental data for the larger economy.

Market data refers specifically to the trading of the securities issued by financial actors whose state of affairs are described by fundamental data. Market data can be divided into data that is available "pre-trade" and "post-trade." Pre-trade market data are bids and offers to buy or sell securities. Post-trade data are records of transactions that have actually taken place. Pre-trade data is an indication of at what prices transactions might occur, and post trade is a record of those transactions. The forms that pre- and post-trade data take vary significantly by asset class. For example, most equities trade on exchanges or venues where quotes are posted, and many countries have disclosure requirements that require dissemination of these quotes. Bonds, which are generally less liquid, often don't have live quotes. Instead, pricing is done on a "request for quote" basis in which the quotes are only valid for a single participant and for a limited time. In markets like these, pre-trade market data takes the form of "indications of interest," rather than a firm offer.

Reference data ties fundamental data and market data to one another and to the economic actor whose affairs or securities transaction they describe. These can be things like stock symbols or other security identifiers. Companies often have multiple subsidiaries which issue debt, so legal entity identifiers help investors be sure which entity actually is the issuer. For market data, transaction identifiers and time stamps help to identify individual transactions or quotes. Reference data gives market participants a common referent when they are talking about fundamental and market data.

Ultimately the different forms of financial data are related. Fundamental data indicates the true state of affairs for economic actors. Market data reflects valuations and transactions in the securities and instruments issued by those economic actors, and thus represents a distillation of the market's views of their fundamental data. Reference data anchors these two kinds of data to actual specific entities using a common identification language and terminology to ensure that market participants are talking about the same thing.

The Evolution of the Ecosystem

This definition and clear categorization of financial data, while useful, is being challenged by some of the forces that are affecting the data world. Three forces are affecting contemporary financial data: the increasing utilization of non-traditional financial data for financial purposes, the spectacular increase in the volume of traditional financial data and a change in the structure of how data is generated and channeled.

The most widely discussed trend is the utilization of non-traditional or "alternative" data in financial services. Recent advances have made it possible for these alternative data sources to add real value to the ecosystem and to clients. Satellite imagery, credit card transaction metadata, data gathered from physical sensors, and data from the public internet systematically gathered are all in this category. In addition to encapsulating a wide range of data types, many alternative data sets are fundamentally different from traditional datasets because of how they are organized or presented.

Traditional financial data falls into the category of "structured" datasets. That is, they have a definite and uniform structure: fixed fields that are the same across every element of a dataset, making them easy to store and search. Market data is a good example of this. A transaction report in equities will possess a fixed number of items: a price, a quantity, a time, and reference data to identify the security and perhaps the venue on which it traded. A thousand transaction reports from the same venue would have the same structure. Techniques for searching, processing and analyzing this kind of data are established and well understood.

Many alternative datasets, however, are unstructured—that is they do not come in standard form and cannot be easily categorized within a relational database. Video and audio files, for example, are unstructured, as is text. Scraping text from the public internet is one of the most popular methods of generating alternative data. Text is unstructured, as it can be of any length and might convey meaning in any number of ways. Indeed, it is possible to derive multiple meanings from the same text or to convey the same idea through any number of textual expressions. This calls for new methods for storing, searching and analyzing text.

The second factor affecting market data is the massive increase in its volume. Changes in law and advances in technology have significantly lowered the barriers to launching and maintaining new trading venues. This has led to an increase in the number of, and consequently competition among, trading venues, thus increasing the number of market data sources. Additionally, regulatory changes that require increased disclosure of both fundamental and market data have increased the volume of data as well. Alternative sources with new structures and increased production of traditional data have combined with changes in data usage to expand the data universe.

Historically, data has been produced and consumed in silos, but these silos are breaking down. Just as alternative sources of data are now being used in finance, different parts of the financial sector increasingly consume data from one another. Take fixed-income exchange-traded funds, for example. ETFs are a type of investment fund that investors can create by assembling baskets of the securities held by the fund or can decompose into those securities by redeeming the fund. ETFs can hold a variety of assets, and those composed of fixed-income securities are very popular. Since these ETFs trade on equity exchanges but are composed of bonds, the price of each affects the other. Thus, anyone trading one needs access to data on the other. This interconnection is playing out with increasing frequency throughout the financial data ecosystem. Therefore, to understand the future of data, it is essential to understand the ecosystem as a whole.

NOTES

NOTES

Cover Illustration: © iStockphoto/ivanastar

The data reported in this document reflect solely the views reported to Greenwich Associates by the research participants. Interviewees may be asked about their use of and demand for financial products and services and about investment practices in relevant financial markets. Greenwich Associates compiles the data received, conducts statistical analysis and reviews for presentation purposes in order to produce the final results. Unless otherwise indicated, any opinions or market observations made are strictly our own.

GREENWICH
DATA | ANALYTICS | INSIGHTS

REFINITIV™
DATA IS JUST
THE BEGINNING