For Institutional Investors Only

Big Data & Investment Management: The Potential to Quantify Traditionally Qualitative Factors

Citi Business Advisory Services

Table of Contents

Key Findings

Big data is a catchphrase for a new way of conducting analysis. Big data principles are being adopted across many industries and in many varieties. However, adoption so far by investment managers has been limited. This may be creating a window of opportunity in the industry.

Investment managers who are able to harness this new approach could potentially create an information edge in their trading that would put them significantly ahead of their peers and allow them to profit from an "information arbitrage" between their expanded models and those of investment managers following more traditional analytic techniques.

Big data technology increases 1) the volume of data that can be incorporated into investment models and 2) the velocity at which that data can be processed.

- Big data is based on the blueprint laid out by **Google** in 2003 around the technology and architecture it developed to handle the massive amounts of data it had to process, store and analyze from retained searches and other applications.
- Big data technologies rely on file-based databases in addition to traditional relational databases. As such, they can store not only structured data, but unstructured data as well. This means that new data sets can be added to a quantitative or systematic model more easily, without a lengthy cleansing, normalization, mapping and upload process.
- Big data technologies also rely on clusters of distributed commodity hardware for processing and on techniques that bounce inquiries from cluster to cluster to utilize any free capacity within the system. This is different than providing point-to-point inquiries to a dedicated server. Because of its distributed nature, big data technologies can process large sets of data at high speeds that facilitate the exploration and testing of new investment hypotheses.

The third facet of the big data phenomenon relates to the variety of data that can now be accessed and the fact that many of these data sets did not exist a few years ago. Parallel to advances in the storage and processing of data has been the development of new types of data being created, primarily due to the growth of the Internet, the advance of social media and the emergence of a new Internet of Things that provides sensory readouts on a huge variety of physical subjects. As these new content sources have developed, there has been a surge in "datafication", which is defined as "the ability to render into data many aspects of the world that have never been quantified before."[1]

With the improved volume, velocity and variety of data inherent in the big data approach, the innovation seen in systematic trading models over the past decade could accelerate. Similarly, a wave of innovation could begin in the quantitative investment space as the differences between what used to represent quantitative versus qualitative research disappear.

- Quantitative fundamental investment researchers could employ big data techniques to expand the number of variables they examine to include data around behaviors, opinions and sensory feedback, areas that were previously only the domain of discretionary fundamental researchers. This could allow for a broader model-based view of what constitutes relative, analogous, superior and inferior value using a new set of data points that are not being incorporated into more traditional investment models. This has the potential to create an information arbitrage between firms that are leveraging big data principles and firms that are not.

[1] "The Rise of Big Data: How It's Changing the Way We Think About the World", Kenneth Neil Cukier and Viktor Mayer-Schoenberger, Foreign Affairs Magazine, May/June 2013 Issue.

- Systematic trading models could process new data inputs at the same volume and speed that current programs use in reviewing price, order and transaction data. Rather than simply selecting trades based on analysis across these traditional data sets, new programs may begin to look for correlations across many inputs and thus prove able to identify new trading patterns that link price activity to non-price related variables. "Multi-factor" systematic programs using this broader data set could realize an information edge that today's multi-system, multi-term and multi-tier systems cannot equal.

- New modeling capabilities linked to the big data approach, such as predictive analytics and machine learning, could change the nature of investment research by creating models that "think" and are able to draw forward-looking conclusions. This could lead to a convergence of quantitative fundamental models that focus on value with systematic trading programs that focus on price. The result could be a new type of automated portfolio management that focuses on "future value" and acts on "likely" events that may not have yet occurred or been announced.

The firms surveyed for this report caution that for most investment managers these changes in approach are still highly aspirational and there are still several obstacles limiting big data adoption.

- **Currently the spectrum of Big Data adoption is broad.** Early adopters are investing heavily in developing a whole technology stack and hiring data scientists to support investment research. Another segment of funds is experimenting with big data by either enlisting big data techniques that extend their existing research capabilities through proofs of concept or by piloting content from third-party providers utilizing big data technology and new data sets. However, based on our survey of investment managers and service providers, most investment firms are not yet focused on big data because they lack the institutional momentum, the skill set and the business case to build out these capabilities in the short-term.

- **Experimentation and usage of big data technology is being driven by the front office and not IT.** Investment firms have been seeking ways to tie big data to alpha generation. In most instances, this effort begins organically. A specific research analyst may put in a request to analyze a new data set to understand its relationship to time-series data leading IT to accommodate the request tactically. Currently, this is not enough to drive wholesale change, but it is beginning to move investment managers into big data adoption. Based on feedback from survey participants, we believe that in 2015 pockets of this type of data analysis will drive a broader array of funds towards a more mature and holistic approach in supporting big data capabilities, similar to current early adopters.

- **Pressure to experiment with and incorporate big data principles into investment research will build because early adopters are already locking up access to semi-private data sets to provide their models with an information edge.** Early adopters are also already using this obscure data in tandem with more readily available data sets, such as social media, government and consumer transaction data. This competency may yield information arbitrage, giving firms that employ big data techniques an advantage over late adopters for some time until these techniques are utilized by more organizations.

- **Efforts to accelerate the adoption of big data principles are being facilitated by a marketplace of third-party providers and data vendors.** This allows a broader swath of investment managers to acquire some basic big data capabilities without full-scale infrastructure and staff investments. Even investment managers who do not acquire teams of data scientists and specialized technology staff will still be able to participate in the evolution of big data in other ways via options discussed later in Section IV.

Other firms surveyed with robust big data programs report that results are not easily obtained.

- **Gaining advantage from big data requires the right set of questions, experimentation and time for patterns to emerge.** Funds that have already experimented with unique data sets have also experienced some failure in obtaining investible insights. Some data sets are not necessarily the right ones for the questions posed due to cultural, regional or other yet-to-be-understood nuances. Funds are spending long periods of time beta testing data as it often takes time for patterns to emerge and because some of the data being investigated is so new that the nature of the data is changing over time. Those that are successful are also not necessarily disclosing where they have found an advantage.
- **There are many integration and cultural challenges that must be understood in bringing new skill sets into the investment management arena.** Many of the new resources coming to investment managers to spur their big data program derive from Internet firms, gaming companies, the military and organizations focused on interpreting and taking advantage of consumer behavior. These resources as well as existing investment management researchers need training to work effectively together.

When it works, however, the benefits of big data are not only being seen in the front office. Key functional areas within funds, such as compliance, are reportedly beginning to rely heavily on big data for emerging use cases such as eDiscovery for trade surveillance or on utilizing outside firms to help standardize compliant uses of social media. Risk teams are looking at running more robust scenario analysis. Marketing teams are looking to examine investor and distribution information to better target capital-raising efforts. With time, investment managers and external service providers may well identify a growing set of non-investment uses for big data that could reduce costs and provide greater operational insight into investment management organizations.

Introduction and Methodology

The ability to mine insights from new types of data in large, complex, unstructured data sets and use new technologies and non-traditional skill sets to probe investment theses is enhancing the way that investment managers perform research. More broadly, the potential presented by big data is also allowing for a new approach to risk management, compliance and other critical investment support functions.

Big data is a larger construct that has been made possible by a convergence of social trends, new data sources, technologies, modes of distribution and merging disciplines. It has yielded a new class of data sets, technologies and an industry of vendors to support it, providing multiple ways for managers of diverse strategies and sizes to take advantage of big data.

To understand the current state of play for investment managers regarding big data, Citi Business Advisory Services partnered with First Derivatives to conduct a survey of industry participants. These interviews were qualitative in nature, focusing on existing use cases, trends, expectations and predictions about where big data principles could take the industry.

Interviews were conducted across a set of investment managers as well as with leading vendors and service providers in the big data space who shared insights about their financial services clients. Where appropriate, some relevant quotes have been included from these interviews to allow readers to experience the statements that helped formulate the views presented in this paper.

An extensive amount of research was also performed, and information and views from many leading thinkers in the big data space as well as from experts on blending teams and creating innovative work environments were utilized. There are numerous footnotes scattered throughout the report and readers are encouraged to refer to these articles and books as a guide if they are interested in more details about any of these topics.

The structure of the paper is as follows:

- **Section I** provides an overview and explanation of what constitutes big data and how it is different from other analytic frameworks available in earlier times.
- **Section II** looks at the new data sets emerging as a result of the "datafication" process and highlights vendors who can be accessed to receive such data.
- **Section III** examines how systematic trading programs and quantitative fundamental programs have evolved and contrasts the current state of the industry with how these investment processes could change as a result of big data principles. These findings are extrapolated to present a possible future where these two types of investment analysis could converge and create a whole new approach of automated portfolio management.
- **Section IV** presents a maturity model that shows how investment management firms can begin their journey toward incorporating big data principles and illustrates how capabilities and management buy-in change as a firm becomes more advanced.

Each of these sections is written with the average front office user in mind. As such, there are only high-level explanations and mentions of the technologies that make up the big data paradigm. The appendix to this paper goes into a much deeper level of detail and is geared toward the IT team within an investment management organization.

Section I: Understanding “Big Data” - New Thresholds of Volume, Velocity and Variety

Big data is a concept that encompasses new technologies and also relies on the melding of new skill sets, analysis techniques and data sets that have only become available in recent years.

To understand how revolutionary the big data concept is, it is instructive to start by contrasting the new approach with what traditional analysis looked like in the pre-big data era.

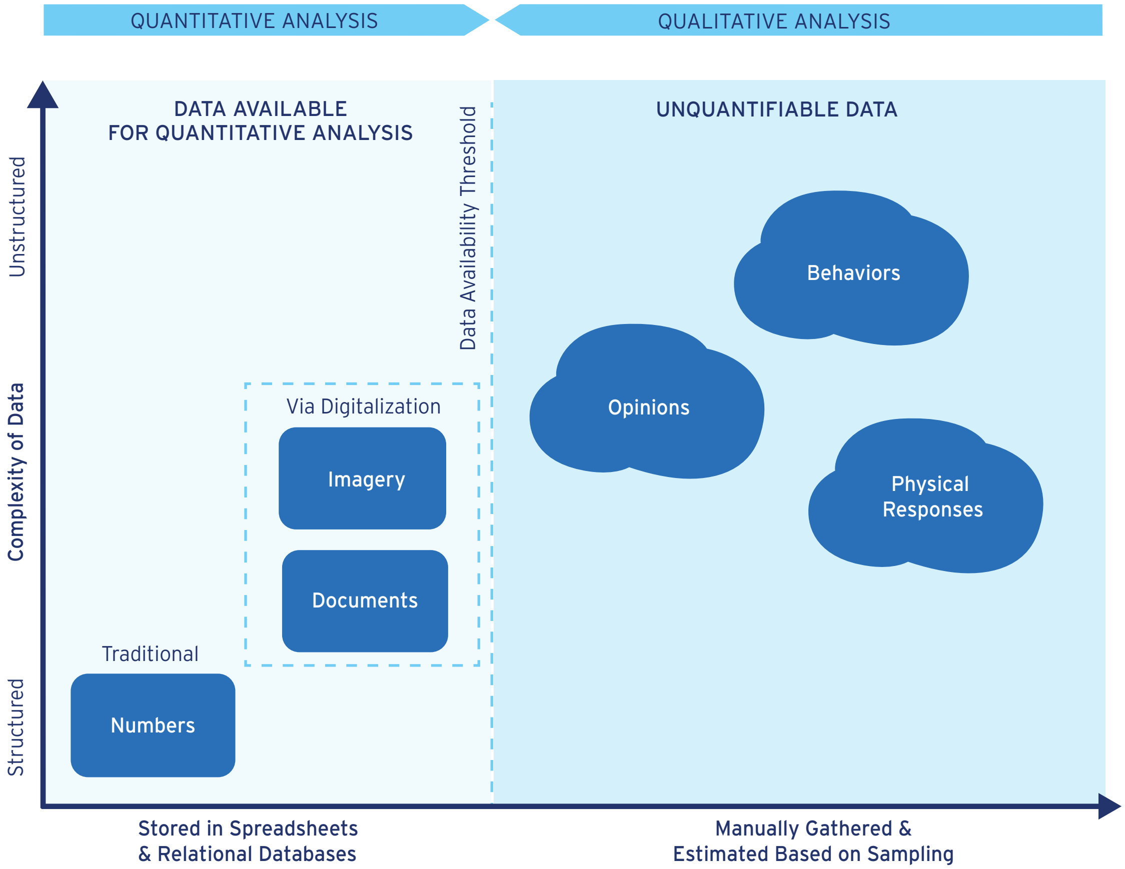
Traditional Quantitative Approaches to Data Analysis

Traditionally, data analysis has implied a quantitative approach, whereby highly structured data is processed in a spreadsheet or a database. **Chart 1** shows that originally this set of data was solely composed of numbers, but as digitalization began in the computer age, teams have been able to take documents and images and convert those data sources into inputs that could also be captured for use in quantitative analysis.

Capturing these new digitized data sets and combining numerical data inputs from several sources typically required substantial normalization and mapping of data fields so that the information could be uniformly accessed. For the past 40 years, relational database management systems (RDBMS) have been processing and storing structured data in both commercial (Oracle, Sybase, MS SQL Server, DB2, Informix) and open source (MySQL) databases.

These databases are characterized as tabular, highly dependent on pre-defined data definitions, and query

Chart 1: Traditional Approach to Data Analysis



Source: Citi Business Advisory Service based ("structured query language" or "SQL"). They require a great deal of upfront work and planning in scaled applications. Even with the emergence of sophisticated standards, adapters, translation mechanisms and indexing strategies for media, they just cannot address all current data needs. Going from proof of concept to full-scale enterprise production applications that use relational databases requires normalization of data streams, which is time-consuming and resource-intensive.

The speed at which data can be evaluated in RDBMS is also highly dependent on the scale of the enterprise's infrastructure; inquiries are processed only as quickly as the speed of the organization's servers will allow. While many groups seek to extend that processing ability by outsourcing their backbone to a managed services provider, the speed at which information can be processed is still limited by the size of the organization's overall platform.

Moreover, it has been nearly impossible for these databases to capture the qualitative inputs that investment teams often use to supplement their quantitative analysis approach. Such qualitative information could include opinions on a company's management team; consumer views of the company's brand; or how the company behaved in pursuing its business goals, such as the cleanliness of their production facilities, the level of controls they had around their supply chain, or how they responded to unexpected manufacturing issues, mechanical breakdowns and employee situations.

Most people gathered their qualitative inputs for analysis through sampling: "Modern sampling is based on the idea that within a certain margin of error, one can infer something about the total population from a small subset, so long as that subset is chosen at random."[2]

Using big data improves upon statistical sampling by making non-traditional types of data available for automated processing and blurs the lines between quantitative and qualitative analysis. While structured data in the form of columns and rows of numbers continues to be processed in traditional relational databases, unstructured data is now available for automated processing within file-based databases, an innovation brought about as a consequence of the Internet age.

Innovation in Database Technology

In October 2003, Google released a research paper on its Google File System (GFS) describing the technology and architecture it developed to handle the massive amounts of data it had to process, store and analyze from retained searches and other applications across its distributed infrastructure.[3] Google developed this system because no commercially available database or centralized server could handle processing the volume and variety of data they were generating and their early custom-built infrastructure had reached its capacity.

In publishing the details of GFS, Google provided the blueprint for developing a distributed data management system that could manage applications and data on clusters of commodity hardware rather than on large single-server systems.

Google's publication of GFS ultimately led Doug Cutting to develop Hadoop in 2006 as part of the Apache Software Foundation's ('AFS') open source stack.[4] Hadoop was adopted and backed by Yahoo and has subsequently been used by developers to process and store large data sets. Because it was open source, Hadoop was embraced and improved on by a large pool of developers and emerging ventures worldwide.

New distributed file-based databases provide several key enhancements over more traditional relational databases.

- Rather than having to transform and normalize data, file-based databases can preserve data in its original format, a convenience that can allow new data sets to be incorporated more quickly and with less dependency on IT resources.
- Since the original formatting can be maintained, there is also no need to decrypt security algorithms and the source material's encryption can instead be kept intact, limiting the ability of unauthorized users to access such data.
- File-based databases are also language neutral. Because of their open source heritage, most use standard processing languages and there are few proprietary technologies that make users dependent on specific providers for enhancements.

[2] "The Rise of Big Data: How It's Changing the Way We Think About the World", Kenneth Neil Cukier and Viktor Mayer-Schoenberger, Foreign Affairs Magazine, May/June 2013 Issue.

[3] "The Google File System", Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, available at http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf.

[4] "HADOOP: Scalable, Flexible Data Storage and Analysis", Michael Olsen, https://blog.cloudera.com/wp-content/uploads/2010/05/Olson_IQT_Quarterly_Spring_2010.pdf.

- Because of the distributed nature of data processing in file-based databases, users can gain leverage in their processing power and realize superior query speeds. This means that there are opportunities to quickly and efficiently analyze huge datasets, including historical data, in almost real-time.

These final two attributes—the ability of file-based databases to handle previously unimagined **volumes** of data and the **velocity** at which distributed infrastructures can process that data—are two of the key components of the big data paradigm.[5] The third attribute is variety, the expansion in the types of data that can be incorporated into analysis.

Datafication and the Explosion in Data Creation

As noted earlier, in the past much qualitative data could not be digitized to fit within relational databases, but increasingly, our ability to digitize that data is improving due to a process called datafication.

Kenneth Neil Cukier and Viktor Mayer-Schoenberger, two of the leading researchers on big data, note that "Big data is also characterized by the ability to render into data many aspects of the world that have never been quantified before; call it "datafication."[6] For example, location has been datafied, first with the invention of longitude and latitude, and more recently with GPS satellite systems. Words are treated as data when computers mine centuries' worth of books. Even friendships and likes have been datafied via **Facebook**."[7]

The impact of this datafication is nothing short of revolutionary. Consider the following. In 2000, only 25% of all the world's stored information was digital. Today, less than 2% of all stored information is non-digital.[8]

File-based databases are the mechanism that is allowing the newly digitized information to be captured and utilized without the need to normalize, map and store that data in relational databases. Yet rapidly changing behaviors are the driving force helping to create ever-expanding quantities and varieties of new data. The following statistics help to drive home this point:

> " The amount of information we've had to provide to date is just a drop in the bucket compared to where this is all headed. It's going to go off the charts."
>
> *– $5.0 - $10.0 Billion AUM Hedge Fund*

- Internet usage has grown to 3.17 billion people online in 2015[9]. More recent user growth has been a consequence of mobile proliferation, extending Internet access to the developing world.
- The amount of available online content is much greater than any person could consume in several lifetimes, even though we know from their activity that Americans are spending a lot of time online. It has been estimated that nearly a thousand minutes of online content is available from social media, electronic commerce and online entertainment for every minute that the average US household has available to consume it.[10] Based on available statistics about online activity in the US, "Every minute we send 204 million emails, generate 1.8 million Facebook likes, send 278,000 Tweets, and up-load 200,000 photos to Facebook."[11]
- It was predicted in 2008 that in the U.S. alone the "Internet of 2015 will be at least 50 times larger than it was in 2006"[12] and estimates are that by 2020 annual data generation will increase 4300%.[13]

These interactions and the content associated with them are providing the fodder that a new breed of skilled experts is able to render into digitized data using a new set of tools to capture, model and analyze that information. This is allowing previously unquantifiable data to now be quantified and that in turn is expanding the threshold of what can be incorporated into data analysis.

[5] "Big Data: Changing the Way Businesses Compete and Operate", Insights on Governance, Risk & Compliance, Ernst & Young, April 2014 available at http://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/$FILE/EY-Insights-on-GRC-Big-data.pdf.

[6] "The Rise of Big Data: How It's Changing the Way We Think About the World", Kenneth Neil Cukier and Viktor Mayer-Schoenberger, Foreign Affairs Magazine, May/June 2013 Issue.

[7-8] Ibid.

[9] "The Statistics Portal", http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/.

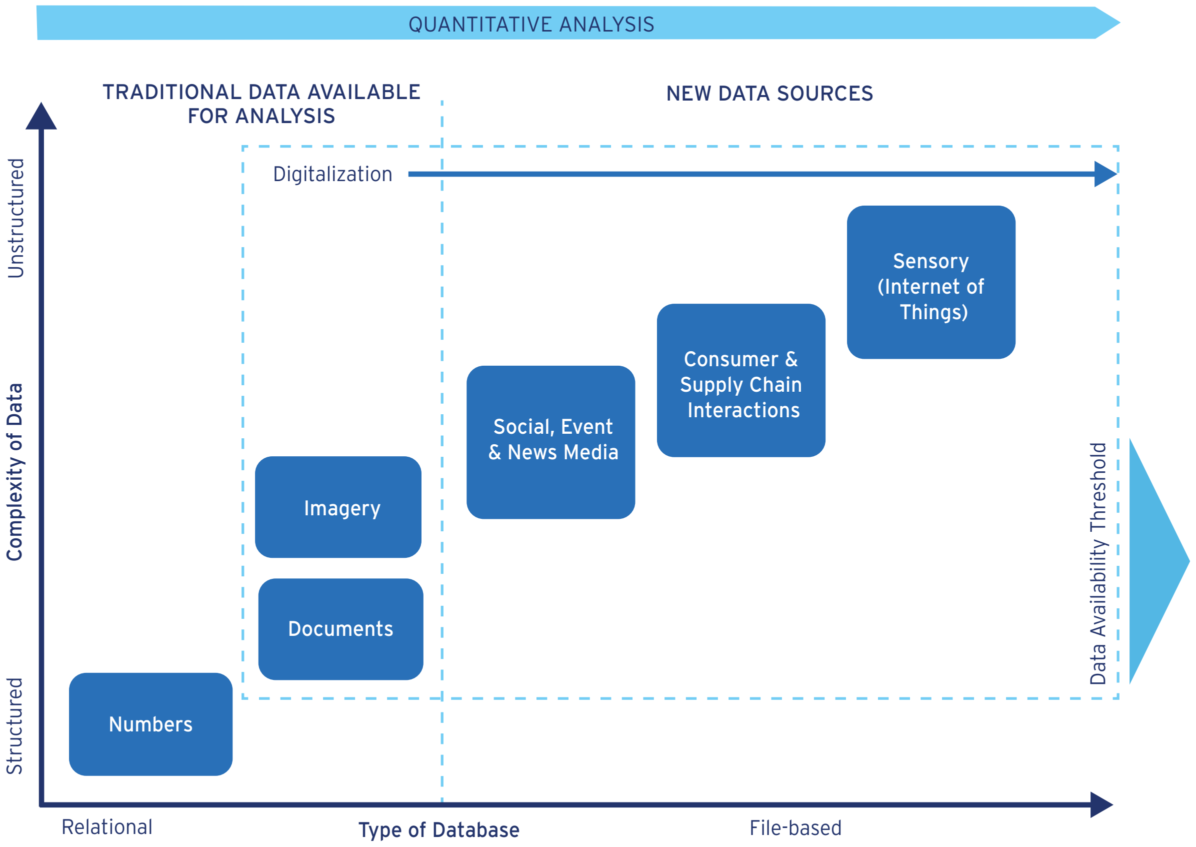
[10] "Tracking the Flow of Information into The Home: An Empirical Assessment of the Digital Revolution in the United States, 1960–2005", W. Russell Neuman, Yong Jin Park, Elliot Panek, International Journal of Communication 6 (2012), 1022–1041, available at http://ijoc.org/index.php/ijoc/article/view/1369/745.

[11] "Big Data: The Eye-Opening Facts Everyone Should Know", Bernard Marr, https://www.linkedin.com/pulse/201340925030713-64875646-big-data-the-eye-opening-facts-everyone-should-know?trk=mp-author-card and Marr's source: http://blog.qmee.com/wp-content/uploads/2013/07/Qmee-Online-In-60-Seconds2.png.

[12] "Estimating the Exaflood, the Impact of Video and Rich Media on the Internet – A 'zettabyte' by 2015?", Bret Swanson & George Gilder, Discovery Institute, January 29, 2008, http://www.discovery.org/a/4428.

[13] "CSC.com, Big Data Just Beginning to Explode", http://www.csc.com/big_data/flxwd/83638

Chart 2: Emerging Approach to Data Analysis



QUANTITATIVE ANALYSIS
TRADITIONAL DATA AVAILABLE FOR ANALYSIS
NEW DATA SOURCES
Digitalization
Unstructured
Complexity of Data
Structured
Sensory (Internet of Things)
Consumer & Supply Chain Interactions
Social, Event & News Media
Imagery
Documents
Numbers
Data Availability Threshold
Relational
Type of Database
File-based

Source: Citi Business Advisory Services

The Emerging Data Analysis Landscape

Chart 2 shows how significantly big data technologies and the datafication of previously unquantifiable data sets are transforming the landscape for data analysis.

More on each of these new types of data and a discussion of the providers in each space will be presented in Section II. The key point to focus on at this time is that the availability of these new data sets has allowed for a vast shift in today's data availability threshold relative to where it was just a few years ago.

The amount of data that can now be incorporated in quantitative analysis is multiples of what it was even as recently as the Global Financial Crisis. This presents many opportunities, but also requires a foundational shift in how users think about data and analysis.

Big data experts Cukier and Mayer-Schoenberger note the following: "Using great volumes of information requires three profound changes in how we approach data. The first is to collect and use a lot of data rather than settle for small amounts or samples as statisticians have done for well over a century. The second is to shed our preference for highly curated and pristine data and instead accept messiness; in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data. The third is, in many instances, to give up our quest to discover the cause of things in return for accepting correlations."[14] These big data principles and how they contrast with traditional quantitative analysis are summarized in Chart 3. In line with the comments above, the approach to data is changing significantly.

Rather than relying on subsets of data and statistical sampling, under the big data approach, the size of the analytic universe (*n*) is approaching the threshold of being able to incorporate all available data into its analysis (*n*=all). The need to manage that data is less, as the standard for data quality moves from pristine to allowing some data sources to be more or less complete and consistent. The goal of analysis is also shifting, as rather than trying to produce an answer

[14] "The Rise of Big Data: How It's Changing the Way We Think About the World", Kenneth Neil Cukier and Viktor Mayer-Schoenberger, Foreign Affairs Magazine, May/June 2013 Issue.

as to "why" something is happening, the outputs of big data analysis instead highlight correlations that set an expectation about "what" will happen.

The following example underscores the importance of this last point. One of the use cases Cukier and Mayer-Schoenberg highlight in their big data research involves Walmart. Walmart saves all of their customer transaction data. By analyzing those transactions relative to many different factors, they were able to identify a pattern under which the sale of the breakfast food Pop-Tarts surged each time the National Weather Service put out a hurricane warning. Having identified the pattern, Walmart moved their Pop-Tart display to the front of their store the next time a hurricane warning was issued and saw sales surge as a result.[15]

Having now created a baseline understanding of what is meant by big data, how it is being enabled and how the principles of data analysis are shifting as a result, the new data sets emerging in the landscape can now be explored with a lens toward how those data sets are being used to extend financial analysis and the analytic tools and new skill sets being used to obtain the best possible value from these data sets can be examined.

Chart 3: Contrasting Approaches of Traditional & Big Data Analysis

	TRADITIONAL DATA ANALYSIS	BIG DATA ANALYSIS
DATA CAPTURE	Sampling (Randomly Selected Subset)	Mass Collection *N*=All
DATA QUALITY	Pristine & Curated	Raw & Varied
ANALYTIC GOAL	Extrapolate a Finding to Describe a Broad Population	Identify Correlations to Predict Outcomes

Source: Citi Business Advisory Services based on findings from Kenneth Neil Cukier and Viktor Mayer-Schoenberger

[15] "Big Data Book Review", by Hiawatha Bray, The Boston Globe, March 5, 2013, https://www.bostonglobe.com/arts/books/2013/03/05/book-review-big-data-viktor-mayer-schonberger-and-kenneth-cukier/T6YC7rNqXHgWowaE1oD8vO/story.html.

Section II: Big Data in Action - New Data Sets and Insights

The variety and volume of new data sets being created through the datafication process are extensive and the opportunities to process vast quantities of such data through the increased velocity afforded by distributed processing are expansive.

Traditional forms of quantitative data are being enhanced now that the content can be more easily stored without extensive normalization and more easily blended into emerging models. New types of data that did not even exist as little as 10 years ago are also being created at a rapid pace by a whole new generation of providers and vendors. In some instances, these data sources are available to any interested purchaser, but in other instances investment firms are looking to lock up semi-proprietary or outright proprietary rights to data usage.

Enhancements to Traditional Forms of Quantifiable Data

Traditional forms of quantifiable data are being expanded to provide bulk access to information.

Time Series Transactional & Government Data:

Investment managers have traditionally utilized time series data around market pricing and curated government data sets around economic and market indicators to inform their analysis.

Many funds have been utilizing massive sets of transaction data in data warehouses and have used online analytical processing (OLAP) technologies and relational databases to manage operational functions, such as position management. Other funds have been handling large volumes and velocities of tick data using more mature in-memory technologies, such as kdb+, to back test and engage in algorithmic trading. They have also mined diverse data sets, such as website screen scrapes, to triangulate and trend user and sales activities on e-commerce sites such as eBay. Specifically, quantitative trading strategies have historically worked with data volume as a core element of their alpha generation process, while global macro strategies have used both data volume and diverse data sets in their analytic process.

File-based databases are able to bring together these existing data sets and allow that information to be expanded with more types of data.

An example of this trend is evident in the movement toward "data-driven government." Nations, states/provinces and municipalities are adopting data-driven governance and policies to make financial, budget, spend, performance and service data more transparent by releasing more documents and archival information to the public. **Havers Analytics**, a set of more than 200 databases from 1,200+ government and private sources, is now being enhanced by an ecosystem of software providers. Companies like **Socrata** are developing platforms to enable distribution of government data to constituents' dashboards for transparency and better engagement. Meanwhile, entrepreneurs and citizen groups are accessing such data to develop apps to improve government services. The growth of a product and service sector enabling the flow of government data at all levels will drive investment research opportunities in regions, industries and instruments based on government activity.

Investment Research Documents:

Some investment managers are aggregating large volumes of historical, written research from sell-side analysts in order to understand relationships and patterns for specific sectors, products and individuals. This can include taking multiple years' worth of content and downloading it for analysis. At the same time, research functions within investment banks have facilitated the process by converting research from digital formats, such as PDF, into data-digestible formats, such as XML.

Satellite Imagery:

Investment managers are also utilizing geospatial data sets from satellite imagery in order to explore visual cues about specific locations. The convergence of less expensive, high-resolution satellite imagery and methods to store the data generated have made it possible to obtain timely imagery and analyze it to inform investment decisions.

In June 2014, Google purchased the company **Skybox**, which produces lower resolution imaging satellites and has on-the-ground technology to store and analyze imagery using Hadoop and standard application languages. Skybox offers analytics on its images relevant to investment strategies, such as "the number of cars in a retailer's parking lot or the size of stockpiles of natural resources in ports",[16] as indicators for how sales may be running for a specific security's financial performance.

Investment managers can purchase these image streams to monitor and assess key industrial facilities in areas as diverse as agriculture, mining or shipping. Understanding the density of crops via imagery of growing locations or measuring shadows on oil storage container images can inform changes in commodities valuations or activities of companies that service these commodities.

An emerging class of data and insight providers, such as **Orbital Insights** and **RS Metrics**, is acquiring bulk satellite imagery and using their own systems and algorithms to perform analysis on these externally produced images to offer investors yet another option. Mainstream press on these providers identifies similar analysis on parking lots as an example use case.[17] However, additional kinds of analysis and their actual effectiveness in beneficial investment decision-making is still emerging.

Data Emerging from Social, Event and News Media

Opinions are no longer something that need to be sampled via focus groups or surveys. Instead, content derived from social, event and news media sites and vendors to interpret and disperse that data are providing data streams that offer insight into the thinking of their myriad users.

Social Media:

Social media sites, such as **Facebook** (1.49bn users Q2 2015), **Twitter** (304mm users Q2 2015) and **LinkedIn** (364mm Q1 2015)[18], are generating massive amounts of user data through posts, Tweets and profiles that can be mined to determine their views, opinions and preferences. Other social media sites, such as **Yelp, Pinterest, YouTube** and **Foursquare**, are developing additional revenue streams around their APIs, where their "check-in" data can be downloaded and analyzed.

Events Data:

A new class of data providers, such as **DataMinr** and **Eagle Alpha**, are parsing Twitter and news feeds in real-time in order to identify global or security-specific events based on "tweet" patterns ahead of and in conjunction with mainstream media, and alerting subscribers of events ahead of wire services. Investors can filter on specific topics, regions, keywords, indices and stocks, receiving alerts for specific events via APIs or directly into Bloomberg terminals, incorporating them into investment models or as commentary for events. Other applications, such as **iSentium's iSense** product, provide a different approach to filtering Twitter by specific US Market stocks, ETFs and some commodities, providing sentiment analysis by looking at reactions to events from mainstream media, bloggers and influencers. In addition to mining insight from event data, Eagle Alpha has expanded its offering to include a marketplace for fundamental asset managers to consume multiple different kinds of alternative data sets, while providing proprietary analytics and research based upon this new breed of investment information.

Another class of media data providers, such as **Minetta-Brook**, utilizes big data and natural language processing to rank and classify news events, blogs, influencers and Twitter feeds around millions of topics. Other established players, such as **RavenPack**, are adopting big data techniques to provide similar news parsing and signaling. In general, subscription costs for these services are more reasonable than building a capability internally and often investment managers are beta testing with a sub-set of users before paying full subscription prices for all their front-office users.

Data on Consumer and Supply Chain Interactions

Behaviors of consumers are also being quantified through a combination of transaction and location services that can follow groups or even unique users via their online and in-store purchases or track point-to-point supply chain movements.

Marketing & Publishers:

The marketing and publisher space occupied by advertising technology companies or marketing firms provides insights to advertisers through their unique access to consumer behavior online. Companies like **AddThis**, **ShareThis** and **Bitly** have been able

[16] http://www.pcworld.idg.com.au/article/547243/google_pays_500_million_buy_satellite_maker_skybox_imaging/.

[17] "Orbital Insight Raised $8.7 Million To Use AI To Analyze Satellite Data", Forbes (March 20, 2015), Alex Knapp, available at http://www.forbes.com/sites/alexknapp/2015/03/20/orbital-insight-raised-8-7-million-to-use-ai-to-analyze-satellite-data/.

[18] The Statistics Portal: http://www.statista.com/statistics/346167/facebook-global-dau/, http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, http://www.statista.com/statistics/274050/quarterly-numbers-of-linkedin-members/.

to aggregate data sets of user behavior through the use of social bookmarking widgets, link sharing and other tools. Other players in this space engage user bases with useful tools and information, offering the potential to provide insight to research teams.

Over 15 million domains place AddThis content engagement tools on their sites and up to 1.9 billion unique users share content on a monthly basis through desktop and mobile devices.[19] In addition to revenue for dashboard reports on user activity from fee-paying websites, AddThis offers advertisers granular insight on consumer behavior. AddThis is additionally looking to monetize their data in raw form via feeds and distribution partners as well as by providing customized dashboards tailored to marketers' needs through their AddThis AI product.

While they do not yet have investment managers as customers, their vast behavioral data set and dashboards may have value to investment managers in providing an understanding of consumer behavior both from an historical perspective and in real-time, having more recently structured their data sets to be consumed and back-tested. AddThis is indicative of the data set potential available once larger data set aggregators figure out how to store, package and feed data for easy consumption.

DataSift is another platform looking to make accessing content easier by unifying and unlocking the meaning of large data sets from social networks, news and blogs. DataSift acts as a single platform to aggregate, query and distribute multiple data sets from their data source partners through streaming and historically through a single API. They provide additional value by normalizing and enriching data and have proprietary classification libraries and taxonomies to make it easier to develop models and enable machine learning via their VEDO rules engine product.

DataSift also offers infrastructure to ensure uptime and failover from source partners. In effect, if an investment manager licenses multiple data sets through DataSift, it can filter on key data from each source company and create a unique and relevant data stream or historical data set for its own purposes. While still early for this market, DataSift already has a handful of investment management clients licensing data through their distribution services and is aggregating marquee data source partners, including **Edgar Online**, **Wikipedia and Wordpress**.

Consumer Transaction Data:

A broader range of funds surveyed is taking advantage of information from financial data aggregators of consumer and small business transaction data, such as **Yodlee** and **Mint**. These providers offer financial institutions and their end customers services to aggregate credit card, debit card, bank account, investment account and mortgage account information. While Yodlee openly markets its analytics services to financial institutions in retail banking and wealth management to optimize customer wallet share by analyzing transactions off-platform, investment managers have been utilizing the enrichment capabilities of these firms to analyze consumer behavior for specific retailers.

The mechanics for obtaining these files and setting up a warehousing database are relatively straightforward, but many of the investment managers surveyed that have done this to date have gone with a legacy approach of normalizing transaction data from different institutions and networks, cleansing the data and then analyzing it in relational databases, because their technical staff and data consumers are comfortable working with SQL and do not have the requisite skills to use file-based storage technologies.

Yet, Yodlee also provides aggregated and standardized reports in addition to data feeds and APIs. For investment managers that do not have the right expertise, programmatic approach or infrastructure, more standard research reports represent an opportunity for investors to leverage the benefits of big data without having to dive in full-scale. Examples such as this illustrate an opportunity for third-party research providers or research teams on the sell-side to utilize big data techniques to provide reports sourced from large data sets.

Location Data and Consumer Behaviors:

Tracking a consumer's location has become a commonplace method of providing information on smart devices. By far the biggest provider of location services has been **Google Maps**, used by more than 1 billion smart phones worldwide.[20]

Whereas Google Maps has gone after the consumer market, the other leader in location services, **Nokia**'s Here, recently sold to a consortium of German automakers **Audi**, **BMW** and **Daimler**, had traditionally focused primarily on offering mapping services to other companies, rather than going directly to consumers.[21] Facebook, Amazon and Yahoo utilize

[19] http://www.addthis.com/.

[20] "Deal Talks for Here Mapping Service Expose Reliance on Location Data," *New York Times*, Mark Scott and Mike Issac, May 12, 2015.

[21] http://www.theguardian.com/business/2015/aug/03/german-car-giants-pay-2bn-for-nokias-here-mapping-service.

Here to be able to overlay addresses on digital maps. Here is also used by **Federal Express** to manage its fleet of delivery trucks worldwide. **Deutsche Telekom**, the German carrier that owns **T-Mobile**, has built smartphone applications with Nokia's services that allow people to share their locations with friends and family through their cellphones and Here additionally holds up to an 80% global market share for built-in car navigation systems, a crucial component for the automakers pursuing driverless car projects.[22]

Foursquare is another consumer-driven location service where users voluntarily "check in" and in turn receive back information, including likes and dislikes, about the retail locations in their immediate vicinity. With approximately 7.5 billion check-ins and more from their companion app Swarm, Foursquare announced in April 2015 that they were introducing Pinpoint, a technology that enables marketers to use Foursquare's data for ad-targeting across mobile devices and the web. This data may eventually become available for other purposes, such as investment research, particularly with Foursquare's API also being used to power location tags for other digital applications like Pinterest.[23]

Another company, **Placed**, has enabled consumer location tracking through a mobile app, whereby a "panel" of consumers, who have granted permission, is continuously tracked. Using changes in location, periodic questions, algorithms and behavioral patterns, Placed can track visits to retailers, frequency and time spent in a store and then link that information to online and off-line promotions, with the possibility of correlating that to sales performance.

This type of data could be used to look at the competitive marketplace, particularly for retailers. Unlike Foursquare, where members actively check in, Placed panel members – numbering over half a million and statistically representative of the overall U.S. population and more granular segments[24] – allow for passive data collection through their movements.

Investment firms surveyed that are investing in equities have shown interest in their data and are actively in beta. However, determining the value of Placed data for an investment manager's purposes is a process, because "the nature of data collected changes over time and time spent with this data is required for patterns to emerge."[25] More on these challenges will be discussed in Section III.

> "We are a nation of individuals. To be able to harvest this data has always been an issue of summation. Now we can do this. Data is going to become like oil. The importance of having it will be amazing because it is going to be the lubricant that makes everything else move."
>
> *– $100 - $500 billion AUM Asset Manager*

Physical Responses Being Tracked via the Internet of Things

The Internet of Things (IoT) is emerging as a major generator of massive data sets from both enterprise and consumer use of remote, connected devices or sensors that receive and transmit information.

Sensory Data:

Convergence of inexpensive sensors, bandwidth, processing, smart phones, wearables, Wi-Fi and expanded Internet addressing (IPv6) are making sensors more ubiquitous. Within industry, sensors are being used to:

- Manage manufacturing
- Monitor agricultural crops
- Signal equipment maintenance
- Monitor gas and oil drilling and pipeline transmission
- Monitor electricity usage
- Manage inventory (RFID)

In each of these areas, the data being relayed by sensors is being captured and stored, creating vast new data sets that can be mined. Some early uses of such data outside the investment arena are already emerging.

Government is utilizing sensors to manage the "Connected City" to improve traffic patterns and to optimize lighting and parking. Smart phones, fitness bands, wearable wireless health monitors and connected devices in the home and in cars are monitoring individual's health status and behavior. Wearable wireless medical device sales are expected to reach more than 100 million devices annually by 2016.[26]

[22] Ibid.

[23] "Foursquare Unleashes Location Data for Cross-Mobile Ad Targeting", Adweek, Christopher Heine, April 14, 2015, available at http://www.adweek.com/news/technology/foursquare-finally-unleashes-location-data-cross-mobile-ad-targeting-164069.

[24] https://www.placed.com/products/insights-use-cases.

[25] Elliot Waldron – Director of Analytics – Placed.

[26] ABIResearch, "Wireless Health and Fitness", 2011 - https://www.abiresearch.com/market-research/product/1005339-wireless-health-and-fitness/.

Companies like **Novartis** and Google are developing devices such as smart contact lenses with sensors to measure glucose in tears for diabetes control.[27] The aggregation of this kind of data has many applications for investment in the healthcare space. However, any devices collecting, transmitting and even aggregating data are subject to HIPAA regulations for mHealth (mobile health) devices, which will require careful due diligence and compliance from investment managers to ensure that data sets and data providers are compliant with these healthcare regulations.

The Information Edge: Semi-Private Data

To some degree, all of the data sets discussed thus far are being created and offered by companies that are looking to build businesses around their supply of such information and are thus looking to attract and support large user bases. Not all such data being pursued by those interested in investment management follow this model, however. There are a growing number of instances where a provider is looking to supply data to a sole or limited set of users at significant price premiums.

Some funds may be seeking semi-private data held by technology companies with robust user communities. This may be viewed as a possible natural progression from "expert" networks, where minable data replaces individuals with closely-held information regarding traffic or trends to color an investment decision process.

While investment managers will have to walk a fine line and make sure that that they are complying with insider trading regulations, where there is no prohibition against mining obtainable data, funds have recognized that these new data sets could possibly provide insight for various investment strategies, investment styles or other operational purposes. Even fundamental research groups in traditional funds are finding value in leveraging data science to reduce workloads for lower-level analysts.

There are numerous examples where semi-private data is already being used:

- The **Financial Times** recently cited an example of a private equity fund "analyzing sales data from the drug-test manufacturer"[28] as an early indicator for employment data.
- Investment funds in the RMBS space are interested in occupancy rates and are using data from providers, such as **Airsage**, to discern occupancy trends by leveraging their aggregated location data derived from anonymous cell phone signals.
- Other funds are seeking data from REITs in order to understand utility usage.
- Where satellite data does not provide imagery of enclosed parking facilities, investors are seeking parking data from real estate holders to triangulate retail activity from parking activity.
- Supply chain analysis and lineage data for large players like **Apple** are being used to triangulate and hedge winners and losers through their massive supplier ecosystems.
- App firms that give away free services, such as email readers, are scrubbing their user base communications for confirmation emails on purchased items and providing these receipts with SKU-level data in bulk to investment managers on a monthly basis

Some investment firms are hiring individuals in the emerging role of Head of Data. This still under-the-radar specialist is not necessarily a technical individual, but someone who understands trading. These individuals are scouring technical trade shows and private companies for minable pockets of information. Private firms are more than willing to provide feeds for data they already own and store for additional revenue.

More recently, regulatory scrutiny of equity markets, particularly in dark pools, has pushed some algorithmic traders to seek advantage in other ways. "It may possibly be the next frontier for funds looking for an investment edge"[29] where the new norm is a "continuous information arbitrage."[30]

> "Individual hedge funds have lined up with individual data providers to suck up information. The leaders in this space are at the point where they can follow specific people around and look at everywhere they go, everything they buy and every preference they show."
>
> *– $100 - $500 Billion AUM Asset Manager*

[27] "Novartis and Google to Work on Smart Contact Lenses", Wall Street Journal, Andrew Morse, July 15, 2014.

[28] "Big data Challenge Perplexes Fund Managers – Madison Marriage", Financial Times, October 12, 2014 - http://www.ft.com/intl/cms/s/0/ffed807c-4fa8-11e4-a0a4-00144feab7de.html?siteedition=intl#axzz3OBbBS9kB.

[29] Ken Cutroneo – Minetta-Brook – Salesperson quote

[30] Elliot Waldron – Director of Analytics - Placed

Section III will now explore what tools and skillsets are needed for this potential information edge and touch on ways in which the industry's current types of quantitative modeling could expand.

Before going into that discussion, however, this section will conclude with Chart 4, a listing of all the different vendors and associated data sets discussed in Section II.

Chart 4: Highlighted Data Providers Organized by Data Set

CONSUMER TRANSACTION DATA	GOVERNMENT	SOCIAL MEDIA	EVENT DATA	LOCATION & CONSUMER BEHAVIOR	SATELLITE IMAGERY	MARKETING & PUBLISHERS
Yodlee	Havers Analytics	Yelp	iSentium	Google Maps	Google Skybox	DataSift
Mint	Socrata	Pinterest	DataMinr	Nokia Here	Orbital Insights	AddThis
		Foursquare	Eagle Alpha	Foursquare Pinpoint		ShareThis
		Facebook	Minetta-Brook	Placed		Bitly
		LinkedIn	Raven Pack	Airsage		
		Twitter				

Source: Citi Business Advisory Services & First Derivatives

Section III: Envisioning the Future of Investment Modeling and Systematic Trading

Investment managers have long sought an edge in their market activities by using data effectively to examine different opportunities. Three types of models currently use this data in an automated manner to realize their market activities. Order analysis systems focus on the timing and release of orders into the markets; technical analysis programs examine chart patterns and indicators to identify, initiate and manage positions; and quantitative fundamental analysis features models that systematically target specific opportunities using model-based criteria.

If the big data paradigm takes hold, there are likely to be new tools, skill sets and models built to take advantage of the revolution in data analysis offered by the big data approach. These advances could transform investment modeling and trading in the coming years.

The volume and velocity of new processing capabilities may offer opportunities to enhance both order and technical analysis programs and result in even more interesting types of pattern recognition and correlations that can produce a differentiated systematic edge. Quantitative fundamental analysis is likely to expand and incorporate an entirely new set of data inputs that have previously been used only in the domain of discretionary trading.

If current trends progress, there could even be a convergence of technical analysis and quantitative fundamental trading models that may result in a completely new type of automated portfolio management. This new type of portfolio management could combine several approaches: 1) machines using proprietary data analysis to select underlying investments; 2) technical programs assessing the likely magnitude of potential price moves based on historical correlations; and 3) the release of orders and management of positions in the market.

As a precursor to showing how much big data principles are likely to change investment management, the evolution of current generation systematic trading models in both the order analysis and technical analysis space will be reviewed first.

Traditional Systematic Trade Modeling

In many ways, systematic trading models are the predecessors of the big data movement. They rely on huge volumes of data and high velocity analysis to operate. Those in the Financial Services arena for the past two decades have seen a foundational shift in the abilities of these models and watched as they have transformed the flow markets. They are, in many senses, a testament to how far those in the capital markets can go in terms of using a combination of analysis and technology to find new and innovative ways of making money.

The only measures that separate these systems from a true big data offering are that they rely on normalized, formatted data that is stored solely in relational databases to run their models and they incorporate only a limited set of price, order and volume data to realize their investment goals.

Traditional systematic trading models have developed in three ways: 1) identifying a **timing edge** by better routing and uptake of order submissions via execution and high frequency trading algorithms; 2) identifying a **trading edge** by pinpointing chart patterns around trends, momentum and mean reversion via market scanning programs; and 3) identifying a **structural edge** by determining differences in the relative value or by recognizing analogous value in varying constituents of a market universe.

A brief understanding of how these models work and the analytic approach they take in determining their trade selection will help to underscore how significant the changes facilitated by new big data opportunities are likely to be in transforming traditional views and attitudes about systematic trading.

Order Analysis via Execution and High Frequency Trading Algorithms:

Execution algorithms began as simple models designed to take advantage of different volume- and momentum-based scenarios. This was because there were notable patterns around when trading volumes or momentum tended to peak or ebb throughout the day for individual markets on specific exchanges.

Trying to link order submissions either to heavy-density volume and momentum periods or to create a volume-weighted or momentum-weighted average to smooth the release of orders throughout the trading day were some of the earliest approaches that used systematic trade modeling to enhance a timing edge.

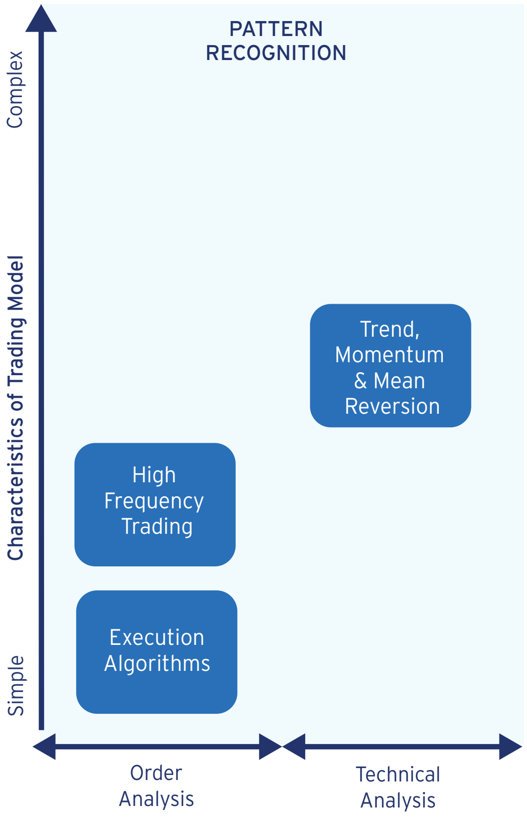
This is a far cry from the high frequency trading (HFT) algorithms that have developed in recent years that look to create and route orders based on perceived market depth, order depth and potential bid-offer inconsistencies across multiple execution venues, with orders often being routed, executed and re-submitted in fractions of a second.

Much has been written both in favor of and against high frequency algorithmic trading in recent years, but for this purpose, the key point to note about both execution and HFT algorithms is that the models built to enable these activities are based entirely around a fairly simple set of standard pricing and order data that is available from key exchanges and execution venues.

As such, the data used to drive the analysis is typically stored in relational, main-memory databases that employ high-powered computing to develop trading algorithms by processing volumes of historical pricing feeds at an extremely high velocity, which then react to market signals in real-time. Though high frequency trading does not employ a variety of data types, the volume and velocity of data used in tandem with a high degree of processing power qualifies this strategy as a forefather of the nascent big data movement within the investment management space.

This is illustrated in Chart 5.

Chart 5: History of Systematic Investment Management



Source: Citi Business Advisory Services

Technical Analysis via Trend, Momentum and Mean Reversion Systems:

Beyond execution and HFT algorithms, there is another category of systematic trading models also built around standard pricing data and geared to respond to specific pricing patterns.

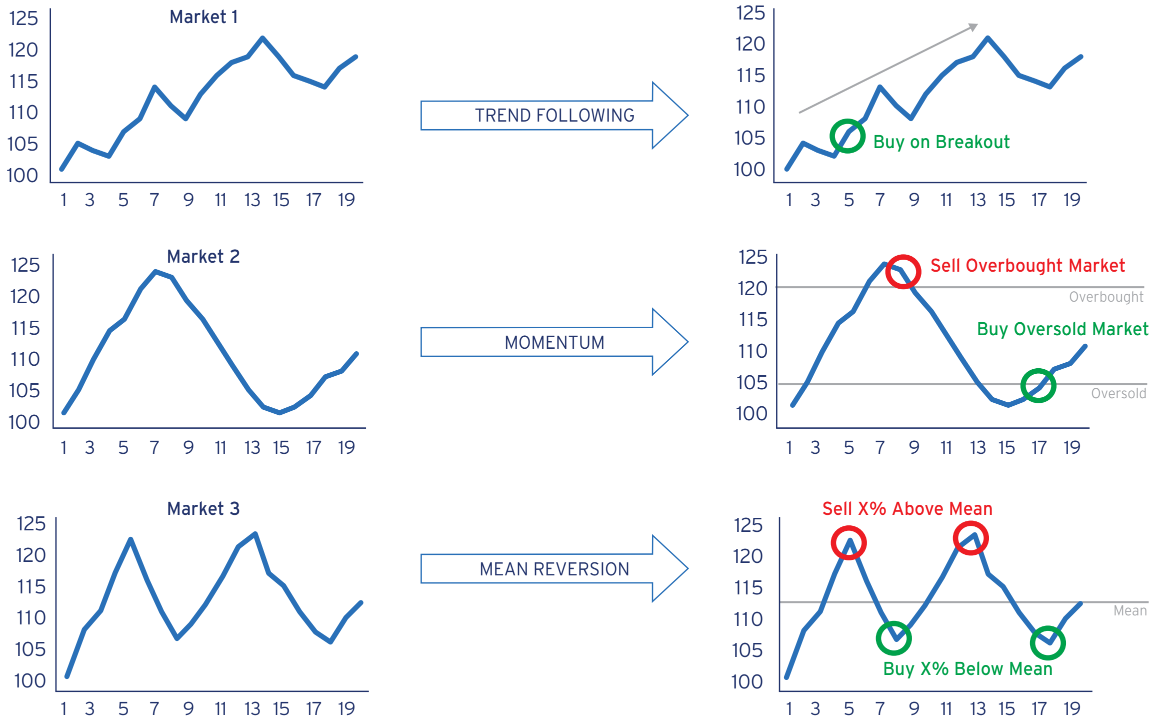
Rather than looking at the best opportunity to execute a trade or capture a bid-offer inconsistency, these models look to identify situations where: 1) prices are trending; 2) prices are showing increased or decreased momentum; or 3) prices have moved beyond a historical mean and are likely to revert back toward that mean. Thus, these models add duration into their assessment and often generate signals that can last across multiple trading cycles. These are the models most typically found in many systematic macro trading models.

While the goal of these programs differs from the execution and HFT programs, the approach underlying the creation of these models is similar in that they rely on pattern recognition around a set of price- and order-related data, they assess data stored in relational databases, they process large volumes of data at high velocity and they generate real-time market signals.

There have been several innovations in these programs in recent years as more and more mathematical talent has been recruited into the investment space. Indeed, some firms surveyed even sponsor scholarships, internships and other programs to create a ready pipeline of such talent.

Originally, many trading systems were designed around daily price charts. Trend-following systems looked at a market over the course of multiple trading days, examining signals such as a cross of the 5-day and 20-day moving averages. Momentum systems would look at breakouts from technical chart patterns or at mathematical indicators based on daily volume and price volatility to indicate whether a market was becoming overbought or oversold. Mean reversion systems would look at a price pattern over many daily sessions (i.e., a seasonal pattern, a cyclical pattern, a

Chart 6: Examples of Key Systematic Trading Approaches



Source: Citi Business Advisory Services

historical price pattern) and compare whether prices in the latest period were showing a below normal or above normal distribution to that pattern in order to signal a trade where the prices could be expected to revert to the mean. These different approaches are illustrated in Chart 6.

As computing capacity became easier to access and market data provision improved in the early 2000s, competition to build new kinds of systematic trading models grew. Innovators began to look at shorter and shorter intra-day time frames (i.e., looking at hourly instead of daily chart patterns) to generate signals and began to develop specialized programs that were flexible and would signal the optimal type of system to use based on the best-fit chart pattern. This enabled them to dynamically select and apply either a trend-following, momentum or a mean-reversion system.

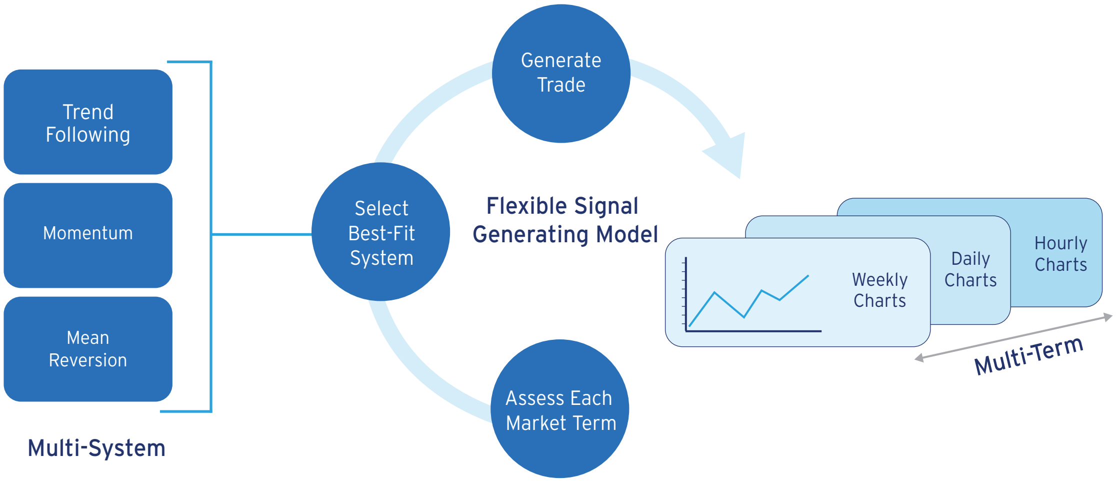
Over time, they began to layer these signal-generating models that move between different types of patterns (multi-system) and would use them simultaneously across a combination of daily charts, hourly charts, minute-by-minute charts etc., (multi-term).

This is illustrated in Chart 7.

Eventually, this impetus to identify as many usable patterns—and profits—from the price action of a single market as possible led to the emergence of multi-tiered trading systems. These systems layer several trading models onto a single market and run those models simultaneously. Signal generating systems are surrounded by transition-monitoring programs that scan for a different set of patterns that may indicate that an earlier pattern is ending. These transition-monitoring programs can turn off certain models just like the signal generating systems can turn on certain models.

In this approach, Signal Generating System 1 identifies a pattern and turns on a trend-following system (System 1A) that may be long a market. System 1A will sustain that long position until System 2, a transition analysis program, indicates that the trending pattern has ended. Meanwhile, System 1 may later identify a secondary pattern, which indicates that the market is moving toward overbought, at which time it may turn on a momentum system (System 1B) that gets short until System 3, a different transition analysis system, determines that the overbought condition has been alleviated. At this time, it will turn off system

Chart 7: Shift to Multi-System, Multi-Term Systematic Trading Models



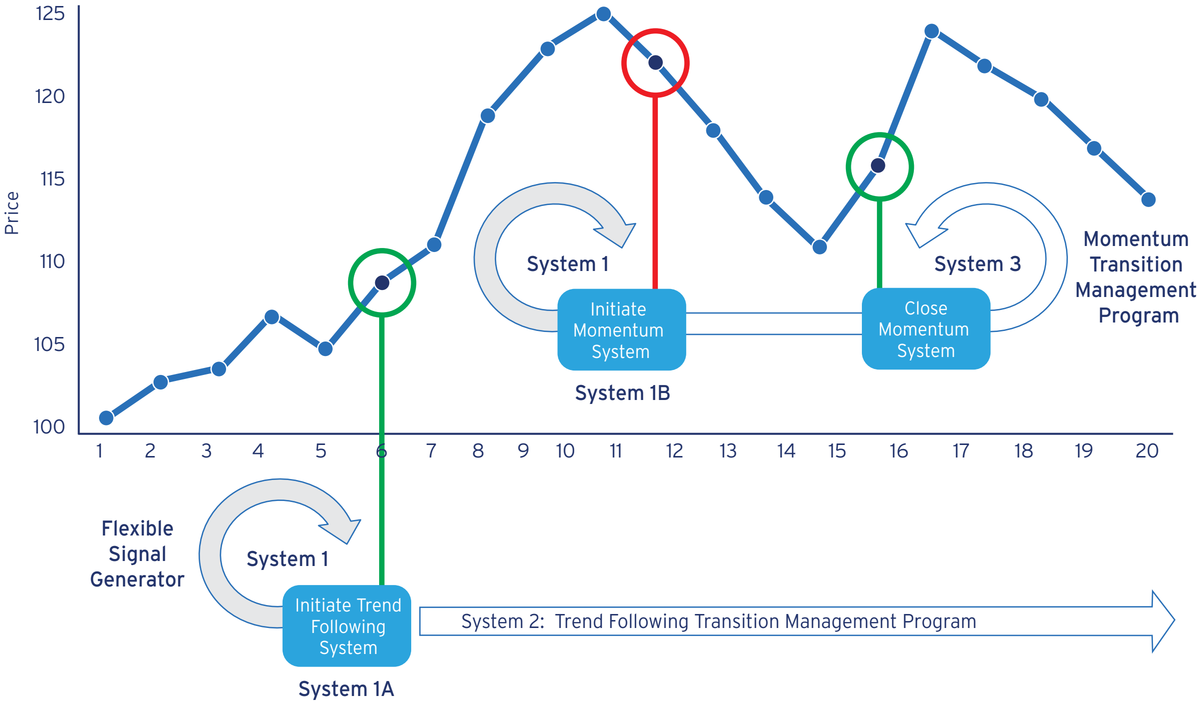
Source: Citi Business Advisory Services

1B and cover the short position. Meanwhile, unless System 2 signals otherwise, the positions put on in response to the trend-following pattern System 1A will remain intact.[31]

Chart 8 illustrates this multi-tier approach.

While the number, type and use of systems in this category of trading models grow more complicated, they still remain focused on a single market and all of these approaches still rely on pattern recognition and on price and order-related data. However, there is an opportunity for these multi-tiered trading systems

Chart 8: Illustration of Multi-Tier Trading with Multiple Systems Running in Same Market



Source: Citi Business Advisory Services

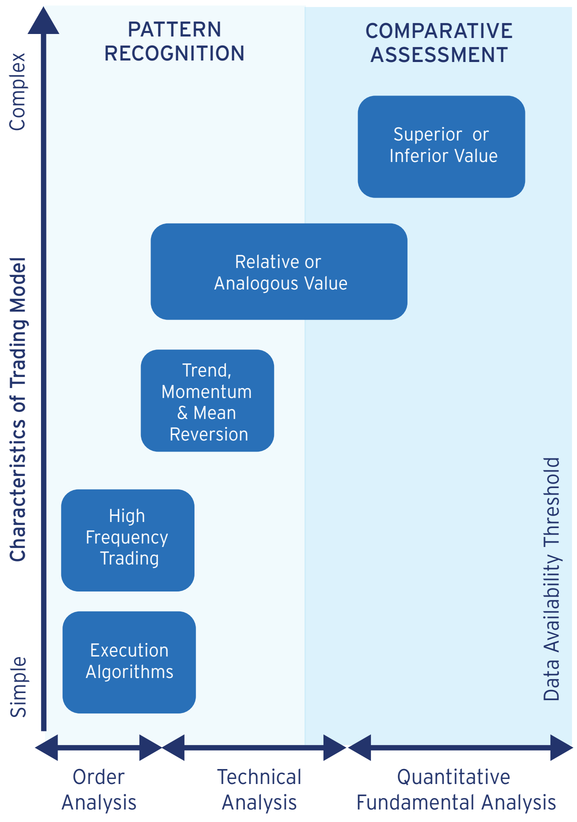
[31] "Moving into the Mainstream: Liquid CTA/Macro Strategies & Their Role in Providing Portfolio Diversification", Citi Business Advisory Services, June 2012.

to be adapted to incorporate some of the new data sets described in the previous section. Additionally, there will possibly be ways to blend the advancements in trade pattern identification and order handling represented by these systems with an increasingly expansive set of quantitative trade models focused on security selection. This potential will be discussed more fully at the end of Section III.

Comparative Assessment to Identify Relative Value and Analogous Trades:

Another type of system looks at patterns across different universes as opposed to within a single market to determine a trading opportunity. Rather than looking at just the price pattern, these systems seek to understand the relationship between two analysis sets and the more sophisticated of these systems will often incorporate another factor beyond price patterns or relationships in an effort to identify a relevant opportunity. As such, they can be considered simple, multi-factor models and they begin to migrate away from just pattern recognition to either a combination of pattern recognition and comparative assessment or to outright comparative assessment. This is illustrated in Chart 9.

Chart 9: History of Systematic Investment Management


PATTERN RECOGNITION
COMPARATIVE ASSESSMENT
Complex
Characteristics of Trading Model
Simple
Superior or Inferior Value
Relative or Analogous Value
Trend, Momentum & Mean Reversion
High Frequency Trading
Execution Algorithms
Data Availability Threshold
Order Analysis
Technical Analysis
Quantitative Fundamental Analysis

Source: Citi Business Advisory Services

The most simple of these systems looks at two trading opportunities within the same market to establish a set of positions.

Trading around seasonality is a common commodities trading technique to identify price patterns and potential turning points within the span of a year. In this approach, a system would look to buy contracts that mature in the peak pricing period and sell contracts that mature during the annual trough.

Energy prices provide a clear example of this opportunity. Europe and the northeastern U.S. are the leading regions for consumption of heating oil and use tends to peak during the winter months, typically in January and February. To have supply in hand for consumption in those months, the oil industry refines and ships heating oil and distributors stock up well in advance. The major seasonal top in heating oil therefore tends to occur in September-October.[32] A system focused on trading this pattern could get long a September 2016 Heating Oil contract and short a March 2017 U.S. heating oil contract and look for a wider than typical carry between those two contracts because of the seasonality of heating oil demand.

Pairs trading is another example of a strategy that looks across different analysis sets, typically within the same market. Models scan for two instruments whose prices tend to move in tandem historically (show a high degree of correlation). When that historical relationship diverges, the model will look to purchase the underpriced instrument and sell the overpriced instrument, with the expectation that the two prices will over time converge back to their historical norm. In a sense, these are multi-instrument mean reversion strategies. Pairs can come in many investment vehicles. Examples include stocks (e.g., Coca-Cola and Pepsi), ETFs (e.g., S&P500 ETF and SPDR DJIA ETF) or even currency pairs (e.g., EUR/USD and USD/CHF).

Some variation of pairs trading is generally considered to be a driver behind most quantitative market neutral, long/short, relative value and statistical arbitrage related systematic trading models.

As was the case with the trend, momentum and mean reversion systems and with the execution and HFT algorithms discussed earlier, systematic seasonal and correlated pairs trading rely on pattern

[32] "Seasonality: Studying Annual Price Cycles," Stanley Dash, TradeStation Technologies Inc., available at https://www.tradestation.com/education/labs/analysis-concepts/seasonality-studying-annual-price-cycles Tradestation.com.

recognition across a standard set of price data. Other models begin to shift the analytic framework, however, from just pattern recognition to incorporate comparative assessment.

An example of systematic trade models that combines pattern recognition and comparative assessment can be found in more advanced arbitrage systems. These models seek to analyze a potential price pattern in one set of data and then express the view about that trade in an unrelated market. The following provides an example of this approach.

A system may analyze the chart pattern of the exchange rate between the U.S. Dollar (USD) and the British Pound (GBP) in order to determine a view on the likely directionality of that currency pair and then use the resulting trend indicator to trigger trades in identical instruments available in both the U.S. and London markets. For example, if the system indicated a trend where USD was likely to rise in value relative to the GBP, the arbitrage system could get long a March 2016 cocoa futures contract on the New York exchange and short an identically sized March 2016 cocoa futures contract on the London exchange. If the currencies move as expected, the U.S. cocoa contract should gain in value relative to the London cocoa contract, thus allowing the system to capture the currency move as expressed by differences in the relative value of cocoa contract prices.

This example is easy to understand because the New York and London exchanges trade identical cocoa futures contracts. Those trading the cocoa market therefore always track the USD/GBP relationship to determine the best location to place hedges.

More complex systems trading in this manner find relationships that are not as obvious. To do so they rely more fully on comparative analysis to identify relevant correlations and less on pattern recognition.

Systematic macro trading firms that already have multi-system, multi-term and multi-tier models running in a given market (e.g., U.S. treasury bills) are now building these more complex systems to be able to amplify their bets by using an analogous market relationship.[33]

The following example illustrates this concept. Through proprietary testing, an analogous trading program may have identified that during the U.S. corn harvest (October-November) there is a 0.25 correlation between the movement of the December corn futures price and the movement of treasury bill prices, and that this correlation has appeared in 85% of back-tested periods over the past 40 years and in 100% of back-tested periods when U.S. inflation, as measured by the CPI, is below 3%. Therefore, this sample system may filter to determine if it is currently October-November and whether the U.S. CPI is under 3%. If the answer to these queries is yes, it will seek to double whatever bets its multi-system, multi-term and multi-tier system may be running in the treasury bill market by putting on 4 December corn futures contracts for each treasury bill contract it is running. Thus, the system is doubling its bet by running positions directly in the treasury bill market and by running an analogous set of positions in the corn market.[34]

The key to these programs is that they are constantly scanning for correlations that they can use to express bets in one market via a secondary market. Nearly all those correlations are still related to price data, but as the example above shows, more factors are starting to be brought into the analytic framework to determine whether there is a go/no-go signal.

In the example above, there was a seasonal filter (the corn harvest period) and there was a conditional filter (inflation below 3% as measured by CPI). Thus these systems can be seen as simple multi-factor models. It is easy to see how the increased **velocity** and **volume** of data being processed due to big data principles can expand the creation of these simple multi-factor models by identifying more and more useful correlations and patterns.

Big data looks to build upon this established practice by also introducing a **variety** of data into the process. The aforementioned satellite imagery can be harnessed to apply spectral analysis to images of corn crops in an effort to determine the level of chlorophyll in the crops. This data becomes yet another input in an established model that correlates to the expected yield of the crop, which in turn may further corroborate the investment thesis.

Quantitative Fundamental Analysis to Identify Superior or Inferior Value:

Multi-factor models are also the underpinning for many quantitative, actively-managed equity and bond trading programs. These multi-factor models expand the analytic framework beyond price to begin to evaluate fundamental factors, including items such as the price-to-earnings ratio, dividend payout and earnings per share in the equity markets; equity-to-debt ratio, volatility and duration in the credit markets; and GDP, inflation and trade gaps in the sovereign debt markets.

[33] "Moving into the Mainstream: Liquid CTA/Macro Strategies & Their Role in Providing Portfolio Diversification", Citi Business Advisory Services, June 2012.
[34] Ibid.

Rather than looking to identify a pair to trade or to find an analogous market to extend a bet, these models look to identify which of the entire universe of instruments they are assessing show superior value and which show inferior value. Different trading strategies can be initiated from these findings.

Benchmarked active long only managers will apply their quantitative fundamental models by overweighting their relevant index components that show superior value and underweighting allocations to components that show inferior value. If a manager were benchmarked against the FTSE 100, they may choose to overweight 25 of those securities, equal-weight 50 of those securities and underweight 25 of those securities in order to outperform the index.

Unconstrained long managers that do not benchmark their funds against a specific index may use the analysis to instead choose to trade a more concentrated portfolio (e.g., only the 25 securities from the FTSE 100 that were overweight in the prior example). Others may choose to look across many different pools of securities to identify optimal opportunities (e.g., sector or geographic rotation).

Managers pursuing new "Smart Beta" strategies will often isolate one or more factors from their multi-factor model and re-weight an index to create direct exposure to just that factor, rather than the capital-weighted measure typically used in index construction.

Each of these strategies reflects a portfolio that is initiated and managed based on the output of their quantitative fundamental model. The output of these models is typically passed to a trading team to implement. Most do not typically initiate trade orders, and thus they cannot be called systematic trading programs. Moreover, their portfolios turn over at a much slower velocity and there may be prolonged periods in which no trade activity occurs and the positions are simply held.

Factors examined in quantitative fundamental analysis are typically derived from data contained in company balance sheets and financial reports or based on data supplied by government or other authorized official sources. Like the price, order and volume data utilized in the prior systematic trading programs highlighted, these strategies also rely on large volumes of tabular data held in relational databases, although there is often not as much pressure to ensure high velocity processing power.

Historically, these data sources and the requirement that they be normalized, curated, stored and queried in relational databases mark the threshold at which quantitative investment management ends. This is noted back in Chart 9 as the "data availability threshold".

Quantitative vs. Discretionary Fundamental Analysis:

Additional insights about the likelihood of an event occurring to change the analytic framework, shifting views about the brand or effectiveness of a company's product or economy, and optimism or pessimism about management or government practices and their impact are also tools used to supplement quantitative fundamental analysis and initiate trades. Yet, these are factors open to **interpretation, not measurement** and as such reside in the world of discretionary, not quantitative fundamental analysis.

Similarly, **predictions** about how an existing set of fundamental influences may change are also a key part of discretionary fundamental analysis. Examples of this type of analysis include ideas about how demand may be impacted due to new research and development efforts and/or the introduction of new products, concerns about shifting demographics, or uncertainty about shipments, deliveries or consumption due to possible weather patterns.

Just as the differences discussed between quantitative and qualitative analysis are disappearing due to the datafication of previously unquantifiable data sets, and the opportunity to extend the analysis set to *n*=all due to the flexibility of file-based databases and the processing capacity of distributed infrastructures, it is evident that the **differences between quantitative and discretionary fundamental trading will also begin to disappear.**

Insights gained from interpretation and from prediction will become "datafied" and help to shift the data availability threshold in investment management. Based on the foregoing discussion, it is expected that a new set of tools and skillsets should help facilitate that shift and increasingly become part of investment management research teams.

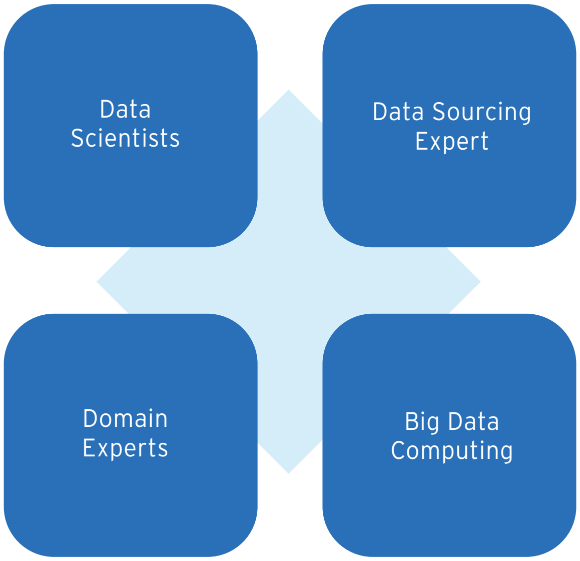
New Tools and Skill Sets Emerge to Advance Big Data Analysis

Quantitative investment research teams have historically been comprised of several key skill sets. Typically, they are led by a portfolio manager with expertise around the investment universe who understands and identifies the various factors making up the investment model. Mathematicians are then employed to assess and draw conclusions from the inputs and outputs of the model and to help think about how to create and evolve the analytic framework. Data modeling experts ensure that the data is normalized, utilized, curated and stored in a manner that allows for consistency in running the model. Computer scientists able to write and run queries and back-tests from the underlying databases enable the rules of the model and help the team identify what tweaks may be needed to the model based on scenario analysis outputs.

In addition to these roles, there are likely to be new members bringing unique skill sets to investment management firms of the future. These new skill sets are listed in Chart 10.

Given the overwhelming variety of new data sources emerging as a result of datafication, investment firms could initially look to add someone whose full-time job is to be constantly **sourcing and securing new data sets**. This individual would be most effective if they understand the drivers of investment decision-making and the supply-demand environment for the target set of investment vehicles. Understanding how data feeds can be created or how information could be systematically collected could also be helpful skillsets, so that the new data sets can be consistently incorporated into the investment model.

Chart 10: New Big Data-Related Skill Sets



Source: Citi Business Advisory Services

Linguistic experts are being leveraged to help shape a new generation of natural language processing tools. The goal is to enable computers to derive meaning from human or natural language inputs. These tools are leveraging the linguistic experts' insights on the identity of language patterns in order to instruct machines to seek similar frameworks in processing documents and other forms of written commentary. This is applicable to the investment management process, as teams can turn research, news reports, Tweets, social network commentary, service reviews and other language-based inputs into value-based signals once they agree the right patterns to scan for and process.

Another shift is likely to be around the types of **domain expertise** required on the team. In traditional quantitative investment management firms, a mathematician has been the main type of domain expert, but many of those interviewed for the survey this year noted that there are other domain experts that are being incorporated. Such domain experts include former military intelligence professionals able to interpret satellite and other images, behaviorists that examine consumer activity to spot emerging trends and shifts in demand, and computer scientists skilled at building predictive analytics and enabling machine learning.

Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. Multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to forecast future probabilities with an acceptable level of reliability.

In predictive modeling, data is collected, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. In a big data context, predictive analytics are used to mine data and forecast trends and social or natural behavior patterns.[35] The form of this that most users are familiar with is when Amazon makes suggestions about the types of books or music you may like based on your previous purchase patterns and on those of others with similar profiles.

[35] Information-Management.com; "Predictive Analytics with Data Mining—How it Works", Eric Segal, February 5, 2005, available at http://www.predictionimpact.com/predictive.analytics.html.

Machine learning tools then take the outputs from natural language processing tools and from predictive analytics to construct algorithms that can autonomously learn and make data-driven predictions or decisions on data. These tools achieve this goal by building a model about likely outcomes as opposed to following strictly static programming instructions. Many view this capability as having evolved from the study of pattern recognition and computational learning theory originally designed to explore artificial intelligence.[36]

Data scientists orchestrate the proper alignment of business questions, statistics, modeling and math to achieve the aims of their employers. Such data scientists are much more multi-dimensional than the typical researcher hired by high frequency trading, quantitative and systematic macro trading firms in past years. Those researchers tended to rely primarily on deep mathematical modeling capabilities. In contrast, data scientists are coming from Internet, gaming and IT firms that have been focused on innovation and consumer behavior.

This is a different type of skill set that ideally will be able to formulate hypotheses that search for correlations that did not even exist until the advent of new data sets. This requires creativity and divergent thinking. The challenge for many organizations may not be in identifying these individuals, but in integrating these resources, as will be discussed in Section IV.

The final skillset which may help bring that edge to fruition will be big data **computing experts** able to bring together the traditional data sets held in relational databases with the new file-based datasets and run that volume of information across a new distributed processing infrastructure to gain speed and variety.

Expansion and Transformation of Quantitative Fundamental Analysis:

Glimpses of the new models that are being built by some investment managers surveyed are beginning to emerge, though nearly all firms involved in this space are very tight-lipped about their efforts. Chart 11 shows how it is anticipated the landscape will change as a result of the insights gained about new models in this survey.

By extrapolating early reports, it is clear that the key to the emerging models will be the expansion in the data availability threshold to incorporate new data sources. These data sets will be derived from the interpretation of existing content and the outputs of predictive analytics. It may not be fully *n*=all, but the extension will allow quantitative modeling to delve much more deeply into a broader sets of hypotheses that were previously only used by discretionary fundamental managers.

> " Who is hiring a chief data officer? That is where the future is going and that is going to be a differentiator."
>
> *– $5.0 - $10.0 Billion AUM Hedge Fund*

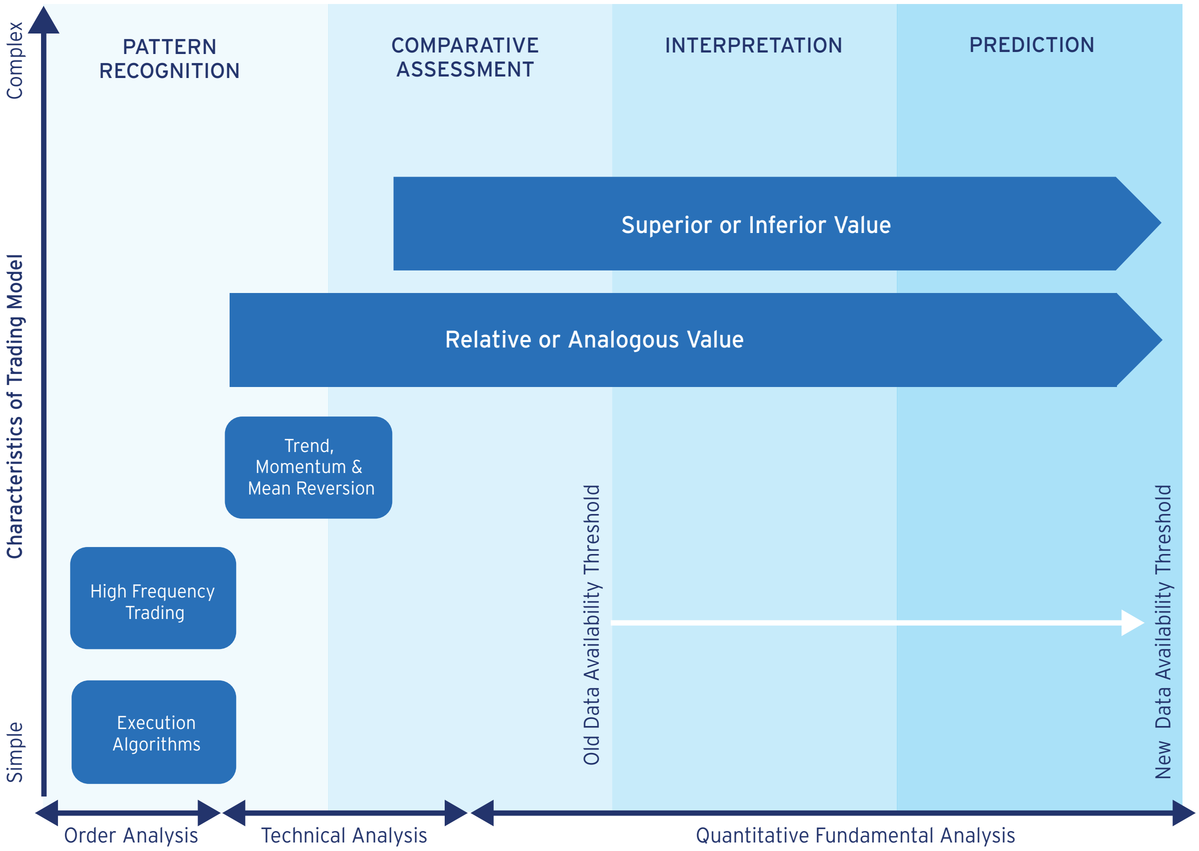
Consider the following example in assessing the Oil & Gas sector. Both a discretionary portfolio manager and a quantitative model are likely to have created a relative ranking of key companies in the sector based on traditional data from balance sheets and financial reports. This basic relative ranking would be as far as most current quantitative models would be able to go.

Meanwhile, a discretionary fundamental portfolio manager might be able to talk to corporate executives and pick up from their body language that they might be excited or nervous about the coming quarter. The discretionary manager could begin to read more about recent activities of the firms and perhaps even visit key offices or facilities to see if they noticed an abnormally busy or quiet level of activity. They could call contacts within the industry and see if they had been picking up any rumors or had been hearing about any interesting new personnel movements.

Through these efforts, the discretionary fundamental manager could determine that the CFO of Company ABC seemed more than normally optimistic in their last visit and that in news coverage of Company ABC there had been several mentions from key executives about how they have been increasing their Research &

[36] "3D Data Management: Controlling Data Volume, Velocity, and Variety", Doug Laney, Application Delivery Strategies, Meta Group, February 6, 2001, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Chart 11: Evolution of Systematic Investment Management



Complex
Characteristics of Trading Model
Simple
PATTERN RECOGNITION
COMPARATIVE ASSESSMENT
INTERPRETATION
PREDICTION
Superior or Inferior Value
Relative or Analogous Value
Trend, Momentum & Mean Reversion
High Frequency Trading
Execution Algorithms
Old Data Availability Threshold
New Data Availability Threshold
Order Analysis
Technical Analysis
Quantitative Fundamental Analysis

Source: Citi Business Advisory Service

Development budget in the past year. Meanwhile, an old retired industry friend mentions that they saw a former protégé who works in new site development for Company ABC at a recent convention and he was exhausted because he was just back from Africa and that this was already the third time he had been there this year. This prompts the discretionary portfolio manager to pull up old reports about undeveloped oil fields owned by Company ABC, two of which are in sub-Saharan Africa. This combination of factors could reasonably allow the discretionary fundamental portfolio manager to predict that Company ABC may soon be announcing that they are opening a new production facility.

Now consider how the datafication of previously interpretive and predictive data can be used to achieve the same outcome in a future-state quantitative investment model.

A linguistic program utilizing sentiment analysis flags that there has been a noticeable shift in the ratio of positive and negative words being used in regard to Company ABC over the past 18 months, with the ratio noticeably trending toward more positive and fewer negative words. In the past four months, the program flags that the pace of change toward positivity is accelerating. In reviewing any increased use of certain phrases, parsing programs identify that company executives in press releases and interviews have used the term "research and development" 182% more in the latest year versus their use of that phrase in the prior 12 months. Secondary words showing increased use include "exploration" and "expansion".

The quantitative model uses the correlation of these words to query public records that list the undeveloped properties owned by Company ABC to determine if any new filings have been made and to understand the locations of all previously filed fields across the globe. The quantitative model could then call up satellite imagery for each location and run an image analysis to determine where there were noticeable changes in topographical features.

The output of this analysis may reveal increased levels of change in both Central America and in sub-Saharan Africa. The quantitative model could then identify individuals that work in Company ABC's new site development team based on an organization chart published in an investor presentation on Google. The model could then reference the Facebook and Instagram accounts of these employees and look at the geo-tagging of all photo postings.

This analysis could identify that across the sample set of 20 individuals, there were 16 photos with a sub-Saharan Africa geo-tag versus only 2 with a Central American geo-tag. This combination of factors would be fed into predictive models that indicate a high likelihood that Company ABC would be opening a new production facility. In fact, the program could go even further and say that based on the progress of buildings on the sub-Saharan African job site, the field is likely to open within the next 90 days.

At this point, both the discretionary fundamental portfolio manager and the quantitative model would establish a long trade in Company ABC based on a likely future event that has not yet been announced.

> "The promise of big data is that we are going to be able to build models at the company and even at the country-basis in almost real-time. The trick will be to tie that together with a sentiment engine. We are not robots, we are emotional people."
>
> *– $100 - $500 Billion AUM Asset Manager*

New Type of "Future Value" Model Could Emerge:

As experience in building these new types of quantitative models progresses and as the ability of systematic trading programs to identify more and more correlations in pricing patterns due to the increased volume and velocity of processing improves, it is likely that the two worlds will collide. The result could be a new type of systematic portfolio management model that bases its trade selection on the likely "future value" of a company. This is illustrated in Chart 12.

These future value models as described by survey participants start with a "future event" output like the one discussed in the Oil & Gas sector example above (i.e., the expectation that Company ABC will be announcing the opening of a new oil field). It would then categorize the type of future event and look for equivalent announcements from the historic records on Company ABC and from its peers. The dates of these announcements would be tagged and charts for the relevant companies in the weeks leading up to and after the event would be examined.

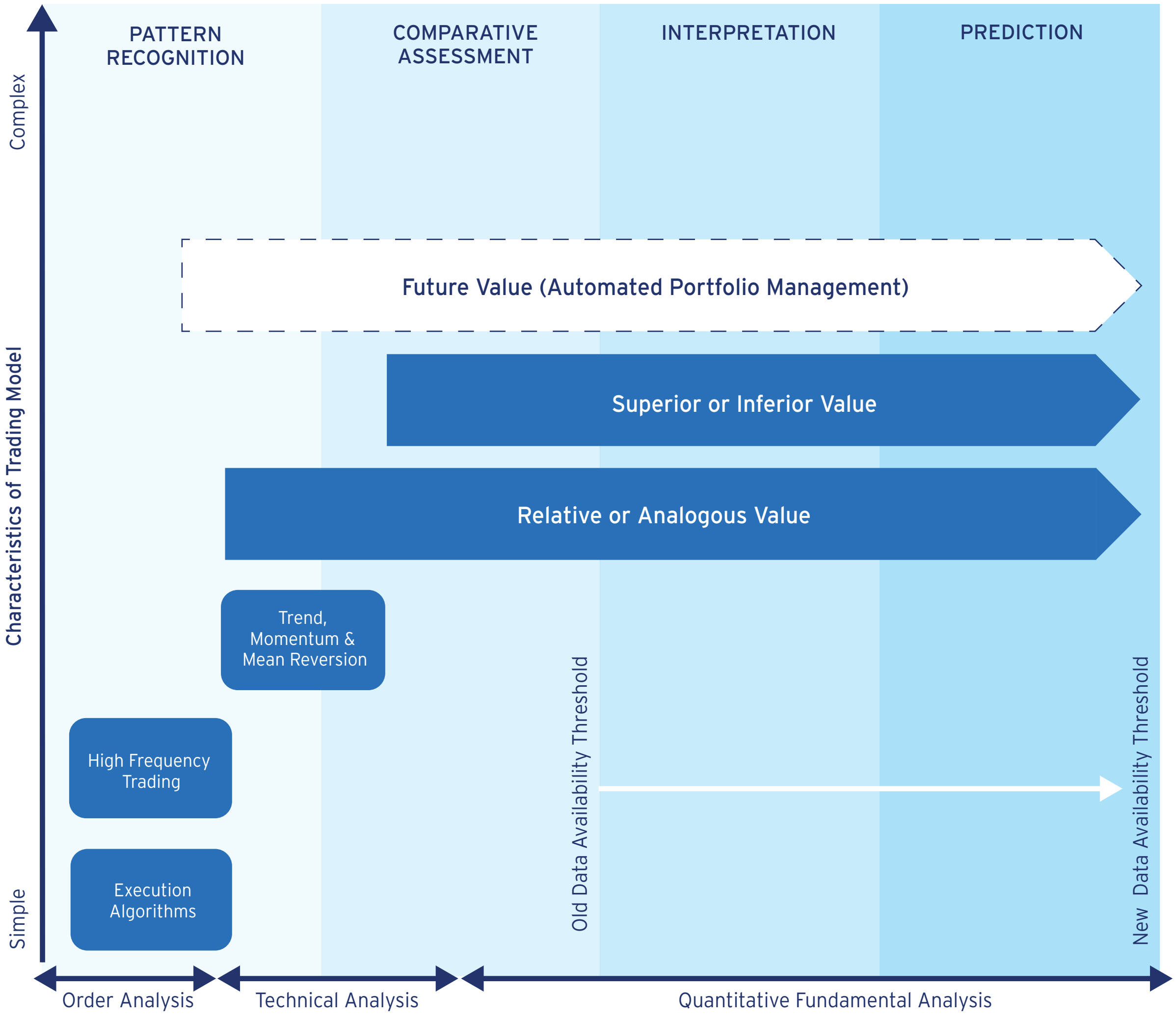
These price pattern programs could quantify the average size and duration of a price jump in response to such news and examine how the size and duration of that move varied based on price patterns in the preceding 90 days and the preceding 180 days. It could examine the latest 90 days of activity and determine which of the historic precedents it most closely parallels. It could then put on a position in anticipation of the future event most in line with that model. Each day, it could reassess the price pattern versus history and determine if the size of the position needs to be increased or decreased based on improving data.

If the event is announced as anticipated, it would close out the position when the expected size and duration of the response to the news has been realized. If the event is not announced within a certain window, it could close out the position or even close out the position earlier than that, if the predictive analysis changes and the likelihood of Company ABC making the announcements erodes.

Other types of "future value" systematic portfolio management programs could be given signals about two companies within the same sector that are both likely to have future events. The programs could thus search historic patterns to calculate how much of a price move and over what duration to anticipate for each individual company. It could then institute a pairs trade in anticipation of a change in the relative value between the two companies. In essence, this would be a forward-looking mean reversion trade.

> "Another way of framing the active versus passive debate is looking at the race between trading being done by algorithms versus trading being done by people."
>
> *– >$500 Billion AUM Asset Manager*

Chart 12: Evolution of Systematic Investment Management


PATTERN RECOGNITION
COMPARATIVE ASSESSMENT
INTERPRETATION
PREDICTION
Characteristics of Trading Model
Complex
Simple
Future Value (Automated Portfolio Management)
Superior or Inferior Value
Relative or Analogous Value
Trend, Momentum & Mean Reversion
High Frequency Trading
Execution Algorithms
Old Data Availability Threshold
New Data Availability Threshold
Order Analysis
Technical Analysis
Quantitative Fundamental Analysis

Source: Citi Business Advisory Services

Other Examples - Big Data Usage:

While the prior example illustrates how a firm might combine multiple types and sources of data for automated analysis in support of their investment thesis, other firms might use only one aspect of this new big data landscape to assist in their investment process.

For example, companies that have an advanced capability in analyzing social media will compare the activity of one company versus another in the same sector to determine the success of competing advertising campaigns.

Other companies might look to satellite images of construction activity in China to gauge the robustness of that economy relative to official government reports.

More commonly, investment managers may look to parse consumer transaction data from a credit card aggregation service in order to forecast retail activity ahead of official survey data.

Some of these methods are effected using the latest in file-based database technology. Others are done using traditional relational databases, but require extensive data preparation in order to allow the data to conform to a sensible tabular / columnar structure. And sometimes these aspects are outsourced to a third-party who is providing the technical expertise, the analysis, or both.

How various investment managers are putting big data techniques into practice and what organizational, operational and technology challenges are emerging as a result will now be explored.

Section IV: Implementing Big Data Programs - From Experimentation to Maturity

Having now established a foundational understanding of big data principles, highlighted the variety of new data sets becoming available due to "datafication", and explored how the ability to process a greater volume of data at a faster velocity could be used to enhance systematic trading programs associated with order analysis and technical analysis, as well as how the addition of a more extensive variety of data could expand the scope and abilities of quantitative fundamental analysis, consideration can be given to how investment managers can look to add big data capabilities to their own firms.

This section lays out a high-level overview of the main data-, people-, process- and technology-related changes required to create an initial capability with regard to big data. An exploration will then be made into how firms have continued to evolve that capability into a full-blown offering. As part of this discussion, the section will lay out at which junctures management must become more fully committed to the big data program, when ownership of the initiative shifts from the IT team to the front office, and what new training and environmental changes might be required to maximize the potential of new big data hires.

For those requiring a more in depth discussion of the technology behind big data, an appendix is included after Section IV in which the specific vendors and their related hosting, hardware and hybrid solutions are discussed. This final appendix is targeted more at the IT professionals within investment management firms to help them initiate their understanding and familiarity with leaders in this emerging space.

Experimentation with Big Data

There are many buzz words that capture the attention of the Financial Services industry and big data has certainly been among these. Chances are that someone within an investment manager has already been discussing big data and making a case for how this could help extend their organization's abilities. Too often, however, the sponsorship for this effort is originating in the IT department, not the front office. Technologists recognize the opportunity that file-based databases and distributed processing capacity represents, but they often struggle to frame that opportunity in terms that the business team can understand.

Building an informed awareness within the senior management team about what constitutes big data and why this marks a significant change in approach from earlier paths is a critical first step. To some extent, this paper itself should help in that education, but finding a successful data set to use to actually show how new information can impact the investment thesis will be a far more compelling and impactful demonstration.

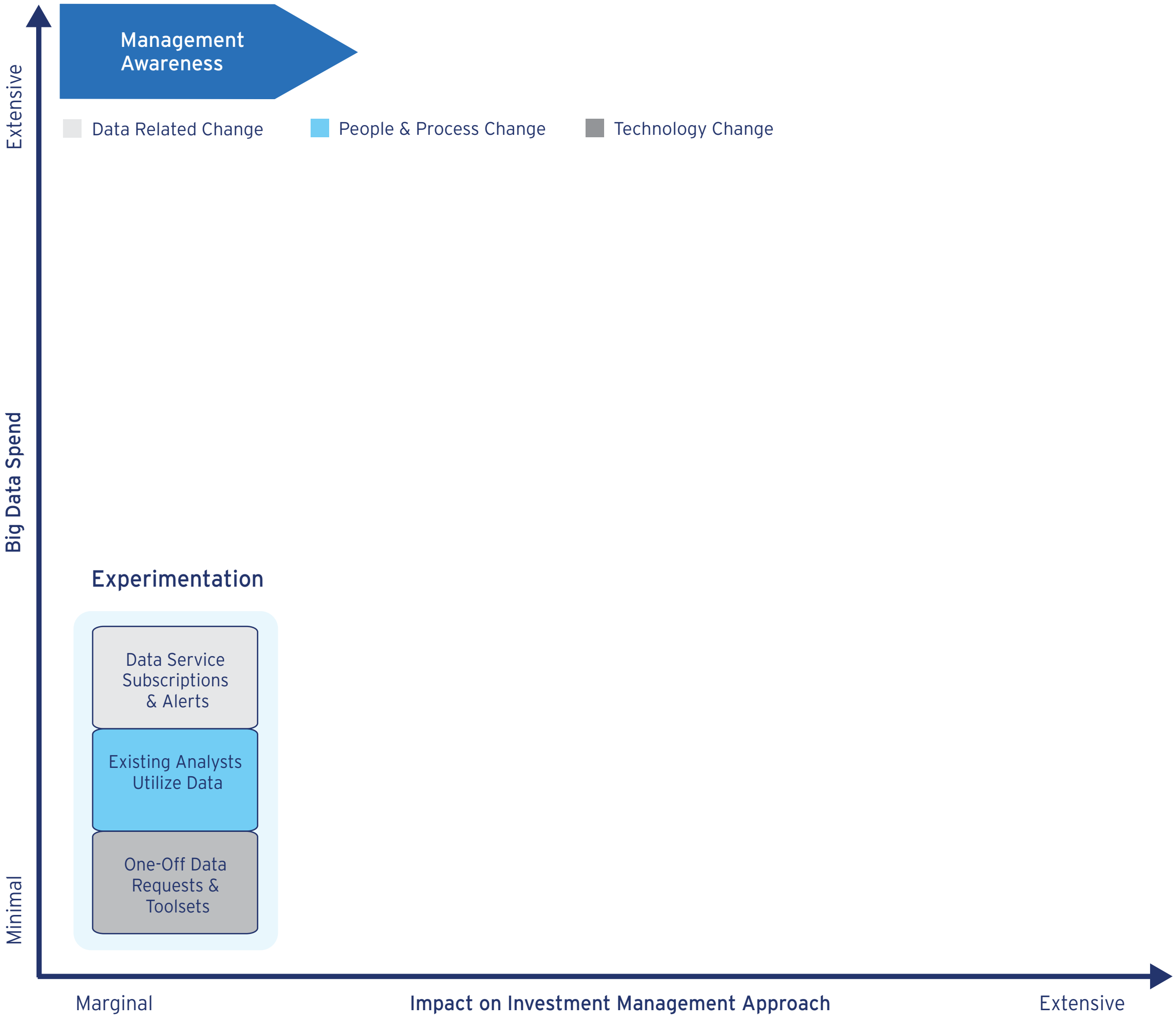
Teaming up IT and a key sponsor from the front office investment team is often the first step. Chart 13 illustrates how this partnership could allow the organization to find a way to experiment with big data while building management awareness.

The lowest barrier to entry to begin experimentation with big data principles would be to subscribe to a new variety of data and look to incorporate that into an existing investment model. As noted in Section II, there are several data subscription services that evaluate emerging data sets and then provide ready-to-consume outputs and alerts around that information. An example would be having your research team receive Google, DataMinr or iSentium alerts through a web browser to provide event color.

Adding this type of new subscription would require only minimal spend and with thoughtful assessment by existing research analysts, could help to show a measurable impact on idea formulation. There would be little to no change to existing systems required in this example, but there would need to be more engagement between the IT team and front office research team to understand how to optimally store and use this information. This approach could also begin to surface some of the new risk and compliance issues associated with such data.

The example above already provides an instructive use case. Information providers, such as DataMinr and Eagle Alpha, add controls around their outbound communications for platforms like Twitter. These controls can be assessed by the organization's

Chart 13: Big Data Maturity Model in Investment Management



Source: Citi Business Advisory Services

compliance and IT departments and help them better define their standards and guidelines around the use of social media.

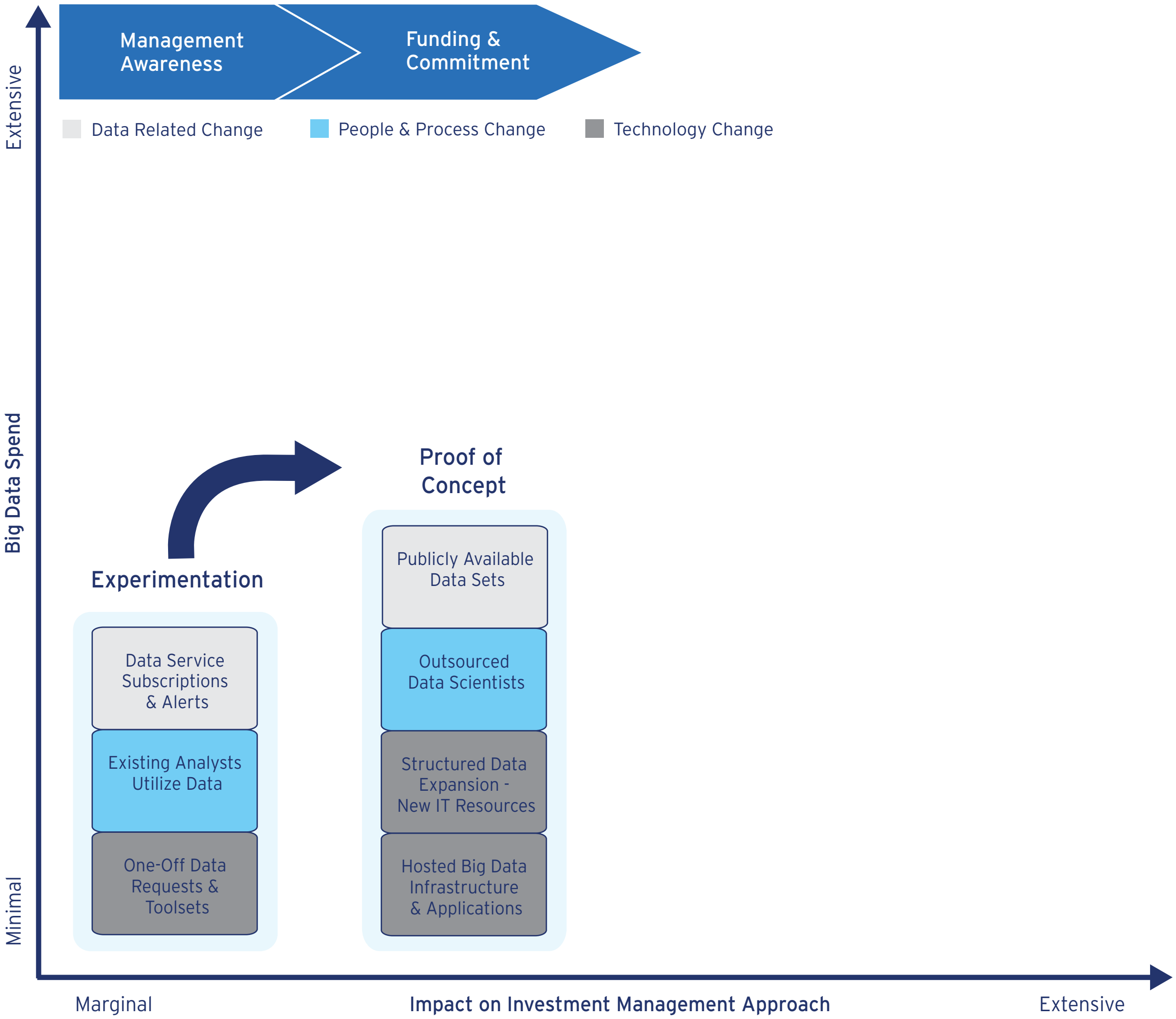
Moreover, IT must be ready to support even the initial analyst delving into a new data set and wanting additional data stores or tools sets to interrogate the data. This may require IT to extend an existing database and business analysts to prepare data for analysis. Compliance as well may insist on the organization storing all inputs used to inform an investment decision, as it is always important to be able to reconstruct how an investment idea was created in case there are ever questions from investors or regulators.

Thus, even this small step will begin to raise awareness about the potential benefits and challenges of big data within the investment and management teams, will lead the organization to consider new compliance and risk questions, and will provide some initial insights for the IT team.

"We've been thinking about how to leverage the unstructured data available through big data for the generation of alpha. The SEC is going to be coming to us and asking us about specific patterns we chose to act upon. Therefore, a lot of the money going to be spent on alpha generation has to be spent on security and monitoring and surveillance."

– $5.0 - $10.0 Billion AUM Hedge Fund

Chart 14: Big Data Maturity Model in Investment Management


Management Awareness
Funding & Commitment
Data Related Change
People & Process Change
Technology Change
Extensive
Big Data Spend
Minimal
Proof of Concept
Publicly Available Data Sets
Outsourced Data Scientists
Structured Data Expansion - New IT Resources
Hosted Big Data Infrastructure & Applications
Experimentation
Data Service Subscriptions & Alerts
Existing Analysts Utilize Data
One-Off Data Requests & Toolsets
Marginal
Impact on Investment Management Approach
Extensive

Source: Citi Business Advisory Services

Moving to Proof of Concept

As awareness of big data builds and early efforts to assess and incorporate new subscription-based data sets progress, there is likely to be a decision made to try to more fully explore how these emerging capabilities could extend the investment management process. At this point, the commitment to a big data program is emerging and there is a proactive choice to begin to establish a budget devoted to the big data proposition.

Establishing a proof of concept around big data can often be a desirable intermediate step before making wholesale investments. This is illustrated in Chart 14.

Rather than solely relying on subscription-based data sets, the organization can begin to expand its variety of data sources in a proof of concept, looking at both structured and unstructured types of data. This middle step in building new capabilities could test a new investment hypothesis by purchasing a specific data set, such as transactional credit card data, or several new data sets, such as satellite imagery and location data.

Because the skill set to incorporate some of these newer data sets into a traditional investment management model may be limited, the organization may choose to look outside the organization to obtain these capabilities through third-party consultants skilled at blending data science into investment strategies. These relationships could be structured on a project basis with defined costs and deliverables.

Firms like **Lucena Labs** are reserving a portion of their investment model business to improving asset managers' existing models or performing bespoke analysis. **Able Alpha** is another type of outsourced consultant that is looking to build systematic trading algorithms using big data principles. Other firms like **Curation** in the United Kingdom are also conducting customized research with big data sets. For example, they have used a picture from Instagram cross-referenced with satellite imagery to verify a key asset in the oil industry. Other tools they use are based on the observation of Twitter feeds from influencers in a given space, traditional public filings and purchased transaction data to triangulate investment signals.[37]

Structured data sets used in the proof of concept are likely to get scrubbed and integrated into the organization's existing IT infrastructure. Most funds still operate with an internal data warehouse that joins end-point systems to support risk, position management and other internal functions. These IT stacks are supported by resources who are knowledgeable in traditional, SQL-based database technologies.

Business analysts from quantitative or global macro funds could be a good source of talent to bring in-house to help facilitate the integration of new structured and unstructured data sets. In many instances, these individuals have traditional SQL-based modeling skills, but they have also made the leap into big data-focused functional languages and have utilized Hadoop environments in their previous roles. This could allow them to work with internal teams and outsourced data scientists to operationalize new investment theses and to scrub and incorporate at least some of the unstructured data for inclusion in the organization's internal data warehouse.

For unstructured data sets that the organization decides to leave in their unstructured state, a firm could look to a cloud-based hosted solution to store such data and provide access to the distributed processing environment that allows for improved volume and velocity of processing. This would include firms like **AWS Redshift** or **Microsoft APS**. These hosted solutions provide multi-tenancy capabilities and thus allow for lower operational costs.

Specialty applications that provide access to big data analysis and visualization techniques via software-as-a-service (SaaS) models could also be leveraged in a low-touch manner during this proof of concept phase. These applications could help investment teams identify, experiment with and understand correlations in visual terms as opposed to simply through quantitative outputs.

By working with a hosted big data solution and a set of SaaS applications, organizations would be able to build understanding within their internal IT teams and allow them to extend their skill set by interacting with these providers. At the same time, it would allow their internal research teams to incorporate new data sets and have environments to test new investment hypotheses without requiring a major overhaul of the organization's existing infrastructure.

In this manner, firms will be able to assess the potential of big data and fit with their investment approach. If the outcome of this proof of concept is positive, the organization may choose to then internalize key capabilities and move toward an early stage big data platform.

> "At the end of the day, I only need to sell one person in my organization on big data—my CEO. This requires a lot of imagination and homework to get up to speed. Right now, other firms are being measured. Allocate a little money to big data and see if it works. You are going in baby steps. For the whole workforce to get there and understand it, it has to be an evolution—not a revolution. Revolutions scare people. Evolutions just happen. You don't even realize it. You just wake up one day and then it's just there."
>
> *– $100 - $500 Billion AUM Asset Manager*

Early Stage Big Data Platforms

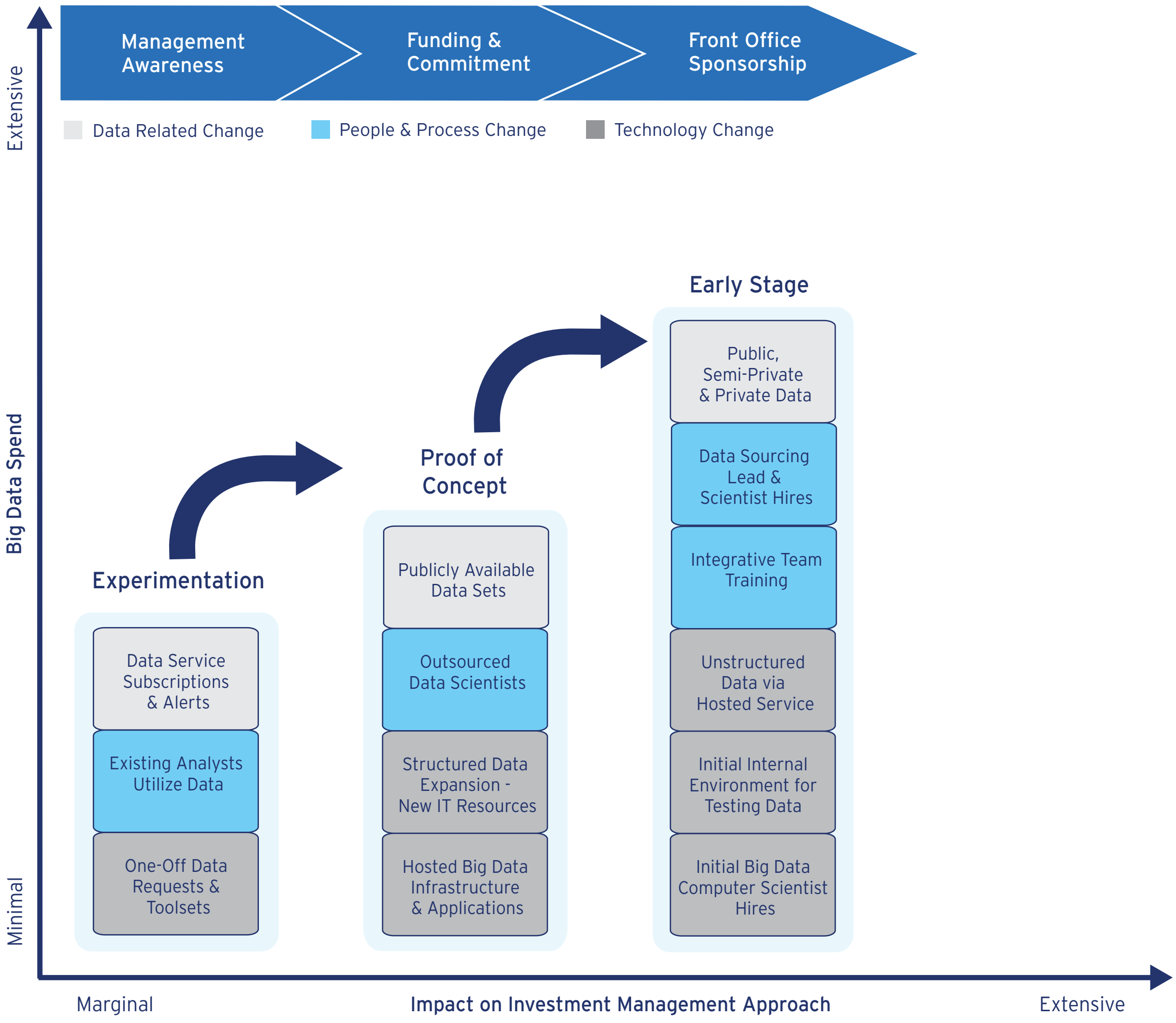
Obtaining front office sponsorship of the big data program is a key next step before organizations should consider spending the money that would be required to build out an internal capability.

While big data techniques could be accretive to compliance, risk, marketing and even operational efforts, the business case for building out these capabilities falls short relative to the amount of organizational change that would be required to properly secure and integrate the resources and technologies that would be required to support an internal big data environment. Tying the build out to an enhanced investment management approach makes these monetary and personnel decisions easier to absorb and ensures that there will be ongoing support as the changes begin to ripple through the organization.

One of the key benefits of choosing to create an internal platform is the ability to have more secrecy around how the investment management process is being enhanced. Indeed, most firms that have gone to this step impose strict controls around divulging

[37] Based on proprietary interview by First Derivatives with Curation.

Chart 15: Big Data Maturity Model in Investment Management


Management Awareness
Funding & Commitment
Front Office Sponsorship
Data Related Change
People & Process Change
Technology Change
Extensive
Big Data Spend
Minimal
Experimentation
Data Service Subscriptions & Alerts
Existing Analysts Utilize Data
One-Off Data Requests & Toolsets
Proof of Concept
Publicly Available Data Sets
Outsourced Data Scientists
Structured Data Expansion - New IT Resources
Hosted Big Data Infrastructure & Applications
Early Stage
Public, Semi-Private & Private Data
Data Sourcing Lead & Scientist Hires
Integrative Team Training
Unstructured Data via Hosted Service
Initial Internal Environment for Testing Data
Initial Big Data Computer Scientist Hires
Marginal
Impact on Investment Management Approach
Extensive

Source: Citi Business Advisory Services

the data sources, the models and the methods by which they formulate their investment hypotheses. Even firms interviewed for this survey required assurances that all examples used in this report would be generalized and based on public examples in order not to endanger their proprietary trading edge.

Secrecy becomes so important because it is typically at this juncture that firms begin to contract for semi-private data sources. Because securing these data sources can so directly impact the investment performance, the responsibility for that function typically falls to a new strategic data head who operates much more as an extension of the investment team as opposed to an extension of the technology team. These and other changes required to build an early stage platform are highlighted in Chart 15.

Firms also begin to hire their own data scientists at this point rather than relying on any outsourcing of this skill set. As noted in Section III, the types of individuals coming into this role are often unfamiliar with the Financial Services world. Many firms participating in the survey note that they have secured data scientists from innovative technology, Internet or gaming firms. Their role on the investment team is to drive new hypotheses that take advantage of previously untapped data sources and modeling techniques.

Integrating these new resources into traditional investment management research teams can be a significant challenge. Traditional investment management teams have worked within a structured environment typically since the earliest days of their careers. As such, they tend to have already adapted to this environment.

Michael Kirton developed the adaption-innovation theory on creating integrated teams. Those steeped in a defined culture often take on adaptive work styles "characterized by working within the given paradigm; these individuals base their approach to problem-solving on understanding the structure of the problem, precision, reliability and conformity."[38] Adaptors seek to solve problems by introducing change that supports the existing system. They hope to provide "better" solutions rather than "different" solutions.[39]

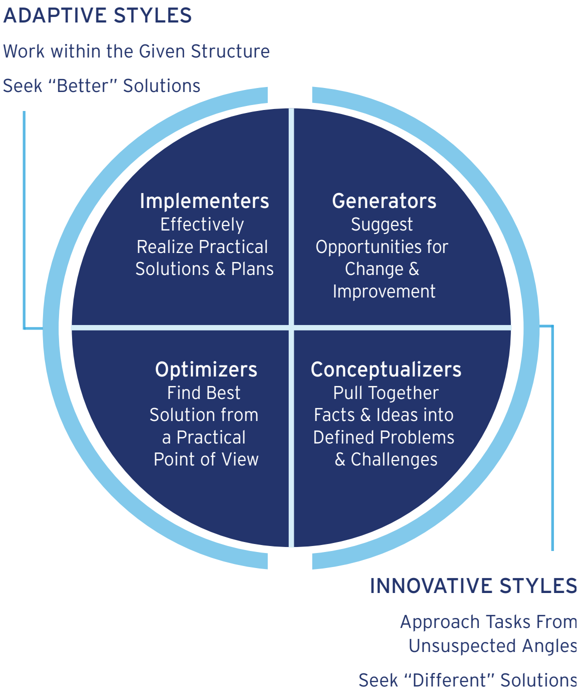
The portfolio managers, analysts and even traders that make up the investment team are all likely to exhibit a high degree of adaptive thinking. Data scientists bring with them a more innovative style that is characterized by "approaching tasks from unsuspected angles, not being limited by the boundaries of the paradigm and being seen as undisciplined."[40] Innovators prefer solving problems with less structure. They seek out different solutions that often leave more adaptive colleagues struggling to recognize and act on such findings. These differences are highlighted in Chart 16.

Having individuals with both adaptive and innovative tendencies within a firm is highly desirable, but often poses a challenge for an organization in terms of integration. Research has shown that a balanced creative team needs a diversity of styles, but that individuals must understand, appreciate and respect each other's styles to make that team function. Moreover, when these teams approach problem solving, they must also understand that within the adaptive-innovative spectrum, there are also sub-roles that people take on and that there must also be an understanding of these roles.

Each individual carries within them certain creative personal styles, often demonstrating one or more styles in their interactions. One of the leading researchers in this area, Min Basadur, has identified four main styles of problem solving.[41]

"Generators" are sensitive to the situational environment and are needed for picking up data and suggesting possible opportunities for change and improvement. "Conceptualizers" are needed to pull together the facts and ideas from the generation phase into well-defined problems and challenges and more clearly developed ideas and concepts for further evaluation. "Optimizers" are needed to find the best solution from a practical point of view.

Chart 16: Creating Well Balanced Teams


ADAPTIVE STYLES
Work within the Given Structure
Seek "Better" Solutions
Implementers
Effectively Realize Practical Solutions & Plans
Generators
Suggest Opportunities for Change & Improvement
Optimizers
Find Best Solution from a Practical Point of View
Conceptualizers
Pull Together Facts & Ideas into Defined Problems & Challenges
INNOVATIVE STYLES
Approach Tasks From Unsuspected Angles
Seek "Different" Solutions

Sources: Citi Business Advisory Services based on work of M. Kirton (1987, 1999) and M. Basadur (1990).

"Implementers" are needed for effectively realizing practical solutions and plans.

Investment management firms looking to bring a new set of innovative thinkers that offers divergent hypotheses and skill sets should be very sensitive to how disruptive these changes can be without adequate integrative training. Leadership and behavioral firms exist to help investment management firms deal with this challenge and lay aside budget to ensure such training may be considered for those pursuing an internalized big data strategy.

From an IT perspective, creating an internal test environment, which begins to integrate structured data from relational databases with unstructured data from file-based databases and is able to run scenarios in a manner that utilizes a distributed set of processors (e.g., a single Hadoop environment), is the first key step in establishing internal capabilities. Until the amount of unstructured data being utilized warrants, it may still be possible to build this test environment while utilizing a hosted service for

[38] "Leading and Managing Creators, Investors and Innovators: The Art, Science and Craft of Fostering Creativity, Triggering Invention and Catalyzing Innovation", edited by Elias G. Carayannis and Jean-Jacques Chanaron, Greenwood Publishing Group, Pages 152-156, 2007.

[39] Ibid.

[40] http://www.kaicentre.com/initiatives.htm.

[41] http://www.basadur.com/howwedoit/TheBasadurProfile/tabid/83/Default.aspx.

file-based data sets and just pull those data sets that are part of the test case into the internal environment.

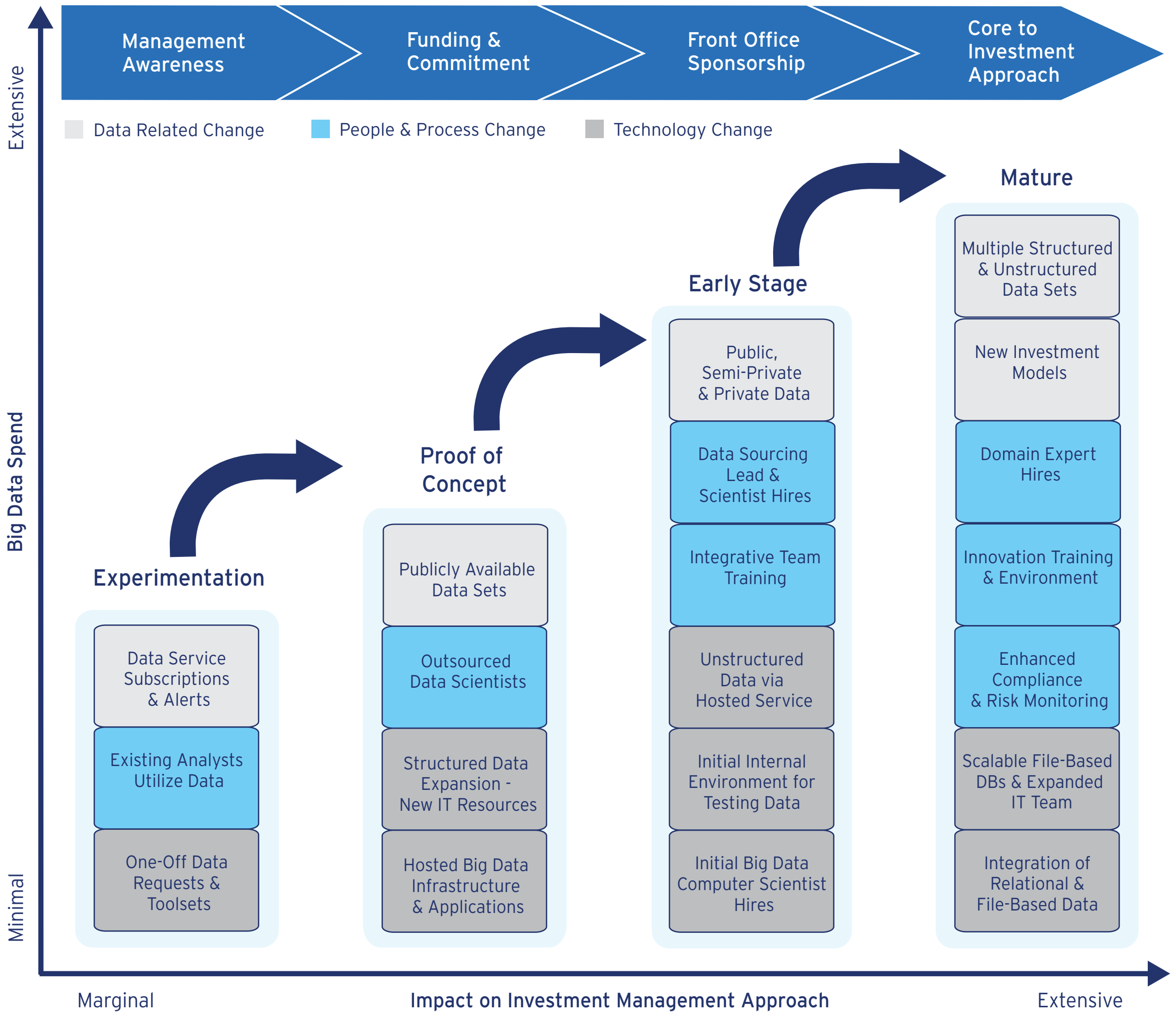
As early stage efforts progress, the organization is likely to have more and more insight into how their legacy investment thesis is evolving. Rather than broad experimentation, specific data sets are likely to emerge that will be regularly incorporated into investment models. This development can eventually push the organization to move beyond early stage infrastructure to a more mature set of capabilities that put big data principles at the core of the investment process.

> "Big data has to be in the business group, not the technology group. In today's world, it is crazy to have a technology group that's focused solely on technology. Every group should be thinking about technology. It is going to be a pervasive part of the business."
>
> *– >$500 Billion AUM Asset Manager*

Mature Big Data Platforms

Survey participants with fully mature big data capabilities are marked by a transformation of the organization at all levels. They describe putting new investment models, processes, resources and technology in place to foster an innovative culture. To provide as conducive an environment as possible, they report noticeable changes in the physical environment of the firm.

Chart 17: Big Data Maturity Model in Investment Management



Management Awareness
Funding & Commitment
Front Office Sponsorship
Core to Investment Approach
Data Related Change
People & Process Change
Technology Change
Experimentation
Data Service Subscriptions & Alerts
Existing Analysts Utilize Data
One-Off Data Requests & Toolsets
Proof of Concept
Publicly Available Data Sets
Outsourced Data Scientists
Structured Data Expansion - New IT Resources
Hosted Big Data Infrastructure & Applications
Early Stage
Public, Semi-Private & Private Data
Data Sourcing Lead & Scientist Hires
Integrative Team Training
Unstructured Data via Hosted Service
Initial Internal Environment for Testing Data
Initial Big Data Computer Scientist Hires
Mature
Multiple Structured & Unstructured Data Sets
New Investment Models
Domain Expert Hires
Innovation Training & Environment
Enhanced Compliance & Risk Monitoring
Scalable File-Based DBs & Expanded IT Team
Integration of Relational & File-Based Data
Extensive
Big Data Spend
Minimal
Marginal
Impact on Investment Management Approach
Extensive

Source: Citi Business Advisory Services

From a data perspective, the few survey participants that have risen to this level report that multiple structured and unstructured data sets from public and semi-private sources are being used as part of the investment process. These inputs have been formalized into new investment models that run in real-time, drawing from scalable instances of the organization's own structured and unstructured databases and taking advantage of the extensive processing power of the organization's own distributed architecture. These and other changes are highlighted in Chart 17.

Upgrades to new investment models are constantly being sought. Some survey participants have been considering hiring domain experts to move beyond the capabilities of commercial applications in terms of trying to build their own filters and analysis tools against the incoming reams and variety of data sources.

The Importance of Innovation Training and an Innovation Environment:

Based on what survey participants report on the integration of big data expertise at their firms, there is an emphasis on training beyond just integration training. Innovation training is a growing field. Most experts point toward needing to develop a dual skill set to achieve true innovation and creative thinking. This relates to employing both "divergent thinking as well as convergent thinking."[42]

Divergent thinking is "characterized by the proposal of a multiplicity of possible solutions in an attempt to determine the one that works. It usually happens in a free-flowing, spontaneous manner where multiple creative ideas are engendered and evaluated. A manifold number of potential solutions are studied in a brief span of time and unconventional connections may be drawn. Once the stage of divergent thinking is complete, information and ideas are structured using convergent thinking."[43]

Convergent thinking focuses on "logic and accuracy and on identifying the known, reapplying techniques and amassing stored information. This strategy is best suited for situations characterized by a readily available answer that just has to be worked out or recalled by way of decision-making strategies. A vital facet of convergent thinking is that it culminates in one best answer, meaning there is no chance for ambiguity. You either have a right answer or a wrong one. This type of thinking is also associated with knowledge (one of the key facets of creativity) as it entails using existing knowledge by way of standard procedures."[44]

Investment teams and their associated risk and compliance support functions are likely to be highly skilled in convergent thinking, but have little experience with divergent thinking. Conversely, the data scientists and domain experts brought into the organization to spur new hypotheses are skilled in divergent thinking, but may have little experience in how to shift from the idea generation to the output phase, which features their ideas actually resulting in trades.

Beyond integrating the different skill sets within the future-state investment management firm, organizations should find ways to help these resources think effectively together. As such, firms may turn to creative thinking and innovation experts. There are firms that can help in that regard.

Strategy consulting firms like **Booz Allen** and **McKinsey and Company** are making affiliations or buying and building out their own innovation teams to focus on the effects that innovation has on business strategy and vice versa. Firms like **IDEO, Continuum, Frog** and **Doblin** come at innovation from a design perspective. Because the design process is geared toward creating actual products, their frameworks may push further than many others that just focus on the thinking part of innovation. Many marketing, public relations and marketing communication agencies have entered the innovation space because they are in the "creative" space and have familiarity with innovation tools. These organizations tend to be light on knowledge transfer, however, as they seek to provide ongoing services rather than teach clients how to innovate.[45] Creativity and training firms like **OVO Innovation, The Creativity Workshop** and **2ThinkNow** offer training around innovation as a process or capability within a firm.[46]

[42] "Idea Generation: Divergent vs. Convergent Thinking", Martin, Entrepreneurial-Insights.com, April 29,2015, available at http://www.entrepreneurial-insights.com/idea-generation-divergent-vs-convergent-thinking/.

[43-44] Ibid.

[45] "The Five Sources of Innovation Consultants", Jeffrey Phillips, Innovateonpurpose.blogspot.com, November 18, 2013, available at http://www.innovationexcellence.com/blog/2014/02/28/the-five-sources-of-innovation-consultants/#sthash.AwyPqvDa.dpuf.

[46] Ibid.

One recommendation that most innovation training firms are likely to make is that the physical environment around a team can have an outsized impact on creativity. There are several aspects of this physical environment to consider.

- The noise level in a firm is an important consideration. A moderate level of noise (70 decibels) versus a low level of ambient noise (50 decibels) is shown to enhance performance on creative tasks whereas a high level of noise (85 decibels) is shown as hurting creativity.[47] Many articles refer to the affinity people have for working in coffee shops because they often fall in the ideal noise range.
- Temperature is also a consideration. A study from Cornell University tested different office temperatures at a large Florida insurance company. They found that when temperatures were low (68 degrees Fahrenheit or 20 degrees Celsius) employees made 44% more mistakes than at optimal room temperature (77 degrees Fahrenheit or 25 degrees Celsius). The study concluded that it was not just being uncomfortable in cold temperatures, but rather that individuals are more distracted. Individuals thus used a lot more of their energy to simply keep warm as opposed to concentrating on creative work.[48]
- Lighting is also something to consider. The Journal of Environmental Psychology looked at the difference in creativity levels in brightly-lit and dimly-lit environments over six studies. The research found that dim lighting helps individuals feel less constrained and free to explore and take risks.[49] Tidiness as well was seen as an inhibitor. A messy space is seen as more conducive to creativity than a tidy one. Studies found that having separate workspaces – a tidy one for structured, task-oriented work and a messy one for creative work – could be optimal for switching between the two modes.[50]

The "messy" workspace can be optimized even more to foster creativity according to several studies. Creative workspaces often involve the build-out of a "thinking room".[51] This thinking room is a place deliberately designed to encourage people to relax and think differently. The goal is to create an environment more like home than work. "Rather than the straight lines, same colors and same desks everywhere" typical of most trading floors, experts recommend that organizations "break things up with different color schemes, curves and texture changes, such as different furniture like armchairs, low tables, carpets or even bean-bag chairs."[52]

The thinking space should also include materials that can be used for random stimulation.[53] This could include a mixed set of magazines, newspapers, books, pictures on the wall and even games to play in breaks and incubation periods. Modeling materials may also help unlock creativity, with experts suggesting materials such "modeling clay (and sculpting tools), wood (balsa is good), wire, pins, clips, glue, nails and so on. A toolkit for cutting, grabbing, bending, hammering can also be helpful."[54] Other suggested supplies to stimulate thinking include "paper of all sizes, from notes to flipchart and rolls of brown paper. Post-it Notes of all different sizes. Whiteboards too (best is full-wall whiteboard). Magnetic whiteboards allow things to be tacked on with magnets."[55] Even video and audio recordings allow teams to capture their creative outputs.

While these may seem like strange concepts to investment managers, there is intellectual study backing the effectiveness of these environmental factors.

Though transforming the investment process has the most potential to spur organizations to adopt big data principles, mature organizations in the big data space can additionally realize many other benefits.

> "We have found it very difficult to integrate some of the new resources we hired from outside the financial services workforce and we have ended up losing a lot of our early hires. We have had to learn the hard way that we need to change our existing way of thinking as much as we have to change our technology capabilities and mix of skill sets."
>
> *– >$10.0 Billion AUM Hedge Fund*

[47] "Impact of the Work Environment: Is Noise Always Bad?", Journal of Consumer Research, Volume 39, No. 4, December 2012, Oxford University Press, available at http://www.jstor.org/stable/10.1086/665048.

[48] "The Science Behind Your Ideal Work Environment," Belle Beth Cooper, Fastcompany.com available at http://www.fastcompany.com/3026715/work-smart/the-science-behind-your-ideal-work-environment.

[49] Ibid.

[50] "Physical Order Produces Healthy Choices, Generosity and Conventionality Whereas Disorder Produces Creativity", Kathleen D. Vohs, Joseph P. Redden, Ryan Rahinel, Carlson School of Management, University of Minnesota, available at http://pss.sagepub.com/content/early/2013/08/01/0956797613480186.abstract.

[51] Creative Workspaces, CreatingMinds.org, available at http://creatingminds.org/articles/creative_workspaces.htm.

[52-55] Ibid.

Other Operational Benefits in Mature Big Data Organizations

Big data has revolutionized the eDiscovery compliance process, whereby suspicious trading activity may be detected by the cross-referencing of trading patterns, news and price movements. The words "anomaly" and "outlier" instantly bring associations with big data, as they imply the inference of sifting through huge amounts of data to determine normal – and non-normal – behavior.

A regulator investigating possible insider trading may try to determine if a trade that produced an outsized profit was a) "out of the norm" for that trader, fund/vehicle or firm, and b) if that trading activity happened ahead of any news that caused c) a significant price movement. As the investment manager looks to police such rogue behavior within their organization, big data technology can alert them by triangulating abnormalities across these three dimensions of trade activity, news and price movement.

Not only will this allow the investment manager to potentially detect and address such behavior as it is happening, but they may also be able to provide evidence to regulators that they have sufficient controls in place so as to demonstrate that any such activity is indeed rogue and therefore not institutional. On the flip side, as organizations move to the type of "future value" models described in Section III, they might use these same capabilities to potentially demonstrate to regulators the signals they used to create their hypotheses and show timestamps and correlations that can help combat any accusations of insider or inappropriate trading.

From the perspective of the risk function, big data techniques can assist in scenario analysis and stress testing, allowing the risk managers to crunch massive amounts of disparate data that can lend itself to predicting uncommon or "tail-risk" events. Strong evidence of the adoption of file-based databases for this risk analysis has yet to emerge, though the storage, distribution and processing power inherent in these new technologies lends itself well to such application.

Core to the marketing effort of firms across various industries has been the attempt to discern customer preferences and behavior. The fund industry is no different from other industries in this regard, and there is evidence that investment management firms have begun to explore how big data techniques can be adapted to help with customer churn analysis and targeting the sales effort. However, as demonstrated in **Boston Consulting Group**'s latest report on the investment management industry[56], there is a gulf between the aspirations for big data adoption within investment management marketing divisions and the implementation of such advanced solutions.

Individually, each of these use cases may be hard to justify for putting out the spend and building the skill set required to succeed in the big data space, but collectively with investment management, the benefits to be gained can be substantial and worthy of the spend, change and commitment required to reach this stage of maturity.

[56] https://www.bcgperspectives.com/content/articles/financial-institutions-growth-global-wealth-2015-winning-the-growth-game/

Conclusion

Big data principles engender a combination of new technologies, new varieties of data and superior processing capabilities that together offer the potential to transform the investment management landscape and the backbone of many investment management organizations.

The goal in producing this paper is to help build awareness of what big data may mean for Citi's and First Derivative's clients. Understanding the opportunities, challenges, vendors and technologies supporting the big data space will arm clients with the inputs needed to help them think through their own big data strategy.

Only a few investment firms are currently advanced in their big data effort, but this paper hopes to show how far this capability can be pushed to expand the way that systematic trading and quantitative fundamental analysis can be enhanced.

This paper is intended to help our clients understand, build and apply a big data program at any stage of their maturity. For those interested in obtaining more information, please reach out to your Citi or First Derivatives sales contact or send a request to Business.Advisory@Citi.com.

For those interested in the specific technologies underlying big data principles, we encourage you to read through the following appendix which explores the offerings from many key providers.

Appendix: Big Data Technology and Vendor Ecosystem

In this section, we would like to delve a bit deeper into the big data technology environment and the emerging universe of vendors providing software and services to the industry.

Depending on capabilities required and willingness to outsource components of the big data value chain, investment managers may utilize the full set or a part of the necessary technologies and services related to big data. Some of these components may already exist within legacy IT environments, especially within more established funds where tools such as data warehouses or data visualization tools have been previously leveraged.

The big data universe consists of internal and third-party data providers/processors, databases of the SQL and Non-SQL variety, application layers, high-level and scripting languages, visualization tools, reporting applications and a variety of service providers – cloud and data centers, systems integrators and specialized vertical service providers, such as **Palantir**, that combine data science expertise, leverage open-source, proprietary applications and niche solutions. This is highlighted in Chart 18.

Big Data Infrastructure and Technology

Hosting and Infrastructure

While big data software is designed for large, distributed networks, investment managers in the survey reported utilizing cloud based infrastructures, particularly public clouds such as Amazon (AWS) and Microsoft (Azure). Privacy and security with regard to proprietary information – strategies, positions and fund investor information – remain key concerns for the majority of PMs and CTOs in the survey, who noted their concern to keep this data and core trading and accounting systems behind management company firewalls.

A select number of PMs reported managing some core functions via hosted application services. They are comfortable with the security in place and want to take advantage of the efficiency and lower operating costs from hosted applications. These funds are just an additional step away from the benefits of big data efficiencies, because these providers leverage big data architectures well suited to multi-tenancy.

A larger segment of fund CTOs reported developing and hosting secondary data sets in the cloud. They described third-party providers leveraging additional AWS services, such as Redshift, to perform data warehouse functions on large sourced data sets.

While not technically a new technology, services like Redshift utilize big data techniques by providing the benefits of massive, distributed and parallel processing for column-based data at a low relative cost ($1,000/year per terabyte), including administration, security, scalability and an ecosystem of integrated third-party products. Especially for POCs, where the data is not mission sensitive, the survey indicated that utilizing cloud capacity is being embraced.

Chart 18: Big Data Infrastructure and Technology

HOSTING AND SERVERS	LANGUAGE / TOOL SETS / APPLICATIONS / PLATFORMS	DATABASES	ANALYSIS AND VISUALIZATION TOOLS
Internal Network / Co-Location / Private Cloud	High-level (Java, C, C++)	Unstructured / File Based / Open Source / Batch (Hadoop)	Visualization tools (Tableau, Qlikview, Spotfire, BO, SAS, Micro Strategy, Platfora)
Public Cloud (AWS, Azure)	Scripted (Python, Perl, PHP)	Unstructured / Document Based / Open Source / Operational (MongDB, HBase)	
	SQL Interface (Hive)	In-Memory / Structured / High-Speed (KX/kdb+, Casasndra, HBase)	
	Complex Data Transformation (Pig)		
	Functional Languages (SCALA, F#, Q)		
	Complex Event Processing (Storm, Spark)		
	Messaging (Kafka)		
	Unstructured Data Discovery (Elasticsearch, Solr)		

Source: First Derivatives

For those investment managers ready to develop in Hadoop, Amazon also offers Amazon Elastic-Map Reduce (EMR), a managed Hadoop framework than enables customers to scale processing across Amazon's cloud infrastructure. It also has products for hosted MongoDB or its proprietary hosted document store, key-value store database DynamoDB. More discussion of Amazon EMR and other hosted Hadoop vendors will be discussed in the Big Data Vendor section of this paper.

Microsoft's Azure hosted service offers an option for hosted Hadoop resources with its HDInsights services, as well as offering additional machine learning, streaming and data management services. More details on Microsoft hosted Hadoop database and service offerings are also described below in the Big Data Vendor section of this paper.

A very select number of funds with deep IT staffs, expertise and a desire to maintain all infrastructure internally are building out their own servers, data centers and even customizing equipment to meet their performance needs.

Storage, Processing and Analysis

As discussed previously, Hadoop is the Open Source database technology that has driven much of the file-based technology adoption currently under way. What makes Hadoop different from previous technologies is its inherent design for large, unstructured data sets within distributed hardware and application environments.[57] Also, the Hadoop stack has matured over the years to provide a set of tools and data processing layers to meet some of the business challenges relevant to investment manager needs, such as:

- SQL-like queries (Hive, Impala, Drill, SparkSql)
- Data transformations (Pig, MapReduce, Spark)
- Streaming data (Storm, SparkStreaming)
- Iterative algorithms and back-testing (Spark)
- Machine learning (Mahout)
- In-memory processing (Spark)
- Messaging (Kafka)
- Full text search and indexing (Solr, Elastic Search)
- Data discovery in large, unstructured data sets
- Storage (HDFS, HBase)

Hadoop and Its Basic Components

The database is comprised of two major components – **HDFS**, the file store component, and **MapReduce**, a job scheduler with other functions that manages processing of data requests.

HDFS stores and decomposes files into smaller blocks of information over multiple servers and disks and constantly checks these servers for disk problems. This inherently makes HDFS scalable and fault-tolerant, because stored data is replicated to another server or storage node if there is an issue. Also, during processing, HDFS load-balances and optimizes processing across multiple stores of the same data. In the simplest explanation, MapReduce enables jobs to be scheduled against stored data in nodes and clusters and split into smaller blocks and even smaller individual field level records. Jobs are also mapped to where the data is stored.

Newer generations of Hadoop decouple MapReduce's resource management and scheduling capabilities from the data processing component, enabling Hadoop to support more varied processing approaches and a broader array of applications. For example, Hadoop clusters can now run interactive querying and streaming data applications simultaneously with MapReduce batch jobs.

Hadoop is flexible in that many coding languages (high-level: Java, C, C++; and scripted: Python, PHP, Perl) can implement a map job, which makes it extremely easy to adopt. Code is also written directly in the database to implement algorithms, so that the code knows where and how the data behaves. These algorithms can utilize simple logic, such as retrievals of single words or numbers, or more complex logic for pattern detection. Results from a job are collected and filtered by a Reducer which sorts and shuffles data and then processes the results, typically writing it back to HDFS. Unlike how data is processed in traditional relational databases, in HDFS the processing is coming to the data rather than the data being queried in the processing.

Hadoop and the tool set around it have matured, such that older criticisms of the technology as being only for batch processing are being addressed. It is still true that not all algorithms work well for Hadoop, but the stack better addresses large scale interactive applications where there are real-time input and output operations. Hadoop is a powerful database than can perform complex algorithms over massive data sets with relative speed and at low cost: "Fault-tolerant, reliable storage under HDFS costs just a few hundred dollars per terabyte."[58]

[57] http://hadoop.apache.org/.

[58] https://www.cloudera.com/content/dam/cloudera/Resources/PDF/Olson_IQT_Quarterly_Spring_2010.pdf.

Ongoing Improvements and Tool Sets Emerging within the Apache / Hadoop Stack

Industry, software firms and open source developers have collaborated more recently on numerous projects within Apache / Hadoop to provide tools sets and performance characteristics that are addressing the collective needs of investment managers to perform more real-time, transactional and varied analysis functions. As discussed above, these projects have made Hadoop into a more mature and enterprise level technology to analyze data sets. The list of the key projects below and the functionality they address relates to analysis relevant to hedge funds:

- **Hive** is the de facto standard for SQL queries over petabytes of data in Hadoop. It is a comprehensive and compliant engine that offers the broadest range of SQL semantics for Hadoop, providing a powerful set of tools for analysts and developers to access Hadoop data. It integrates with all existing business intelligence and visualization tools through ODBC and JDBC connections, making it easier for hedge fund research and IT teams to leverage existing coding skill sets.

- **Hbase** is a non-relational (NoSQL) database that runs on top of HDFS. It is columnar and provides fault-tolerant storage and quick access to large quantities of sparse data. It also adds transactional capabilities to Hadoop, allowing users to quickly conduct updates, inserts and deletes. Hbase has a SQL layer called Phoenix for easier access that enables analysis of structured data, such as price and transactions.

- **Pig** allows you to write complex business transformations using a simple scripting language. Pig Latin (the language) defines a set of transformations on a data set such as aggregate, join and sort. Pig Latin is sometimes extended using UDFs (User Defined Functions), which can be written in Java or a scripting language and can then be called directly from Pig Latin.

- **Storm** is a distributed real-time computation system for processing fast, large streams of data. Storm adds reliable real-time data processing capabilities with low-latency dashboards and data feeds, security alerts and operational enhancements integrated with other Hadoop applications in a cluster.

- **Spark** allows data scientists and quants to effectively implement iterative algorithms and back test for advanced analytics. It is an alternative to running some discreet data science workloads.

- **Mahout** is a library of scalable machine-learning algorithms, implemented on top of Hadoop and using the MapReduce paradigm. Mahout has been used extensively for recommendation engines based on previous outcomes from large user data sets.

- **Kafka** is a fast, scalable, durable and fault-tolerant publish-subscribe messaging system. Kafka is often used in place of traditional message brokers like JMS and AMQP because of its higher throughput, replication and fault tolerance. Applications for this could be in messaging for algorithmic trading.

- **Solr** is the open source platform for searches of data stored in HDFS. Solr powers the search and navigation features of many of the world's largest Internet sites, enabling powerful full-text search and near real-time indexing. It handles search for tabular, text, geo-location or sensory data.

- **Elastic Search** is a flexible and powerful open source, distributed, real-time search and analytics engine. It is architected for use in distributed environments where reliability and scalability are critical. It enables full-text search.

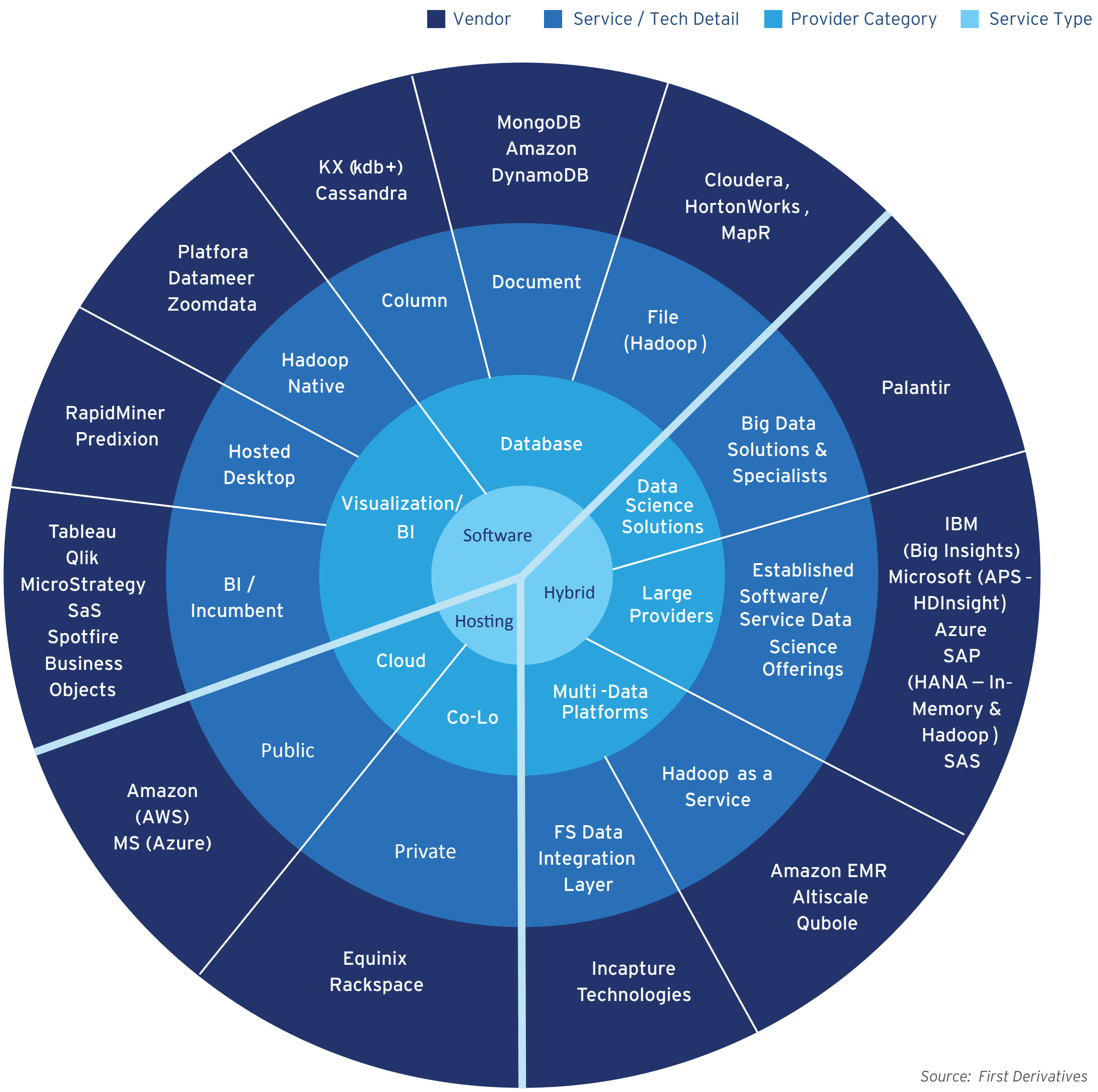
Visualization and BI

The Hadoop stack easily integrates with the common set of business intelligence visualization desktop and enterprise solutions on the market. Tools such as Tableau and Qlikview sit above applications and file-based databases to present analysis for user consumption.

In this regard, analysts in the front office can interface with data in similar ways, using tools with which they are already familiar. There is still work involved in transforming Hadoop data into structured data; analysts are still dependent on specialized coding skills with Apache tools, such as Hive and Pig, to transform and structure the data. Only then can the data be ingested by BI tools and their visualization components.

Additionally, there is a new class of tools for business analysts to interface with data sets more native to Hadoop. More than just a visualization tool or an existing BI suite, platforms like Platfora, Datameer and Zoomdata can work with raw data in Hadoop and reduce some of the data transformation workflow required via ETL and intermediate open source applications (e.g., Hive and Pig) to perform SQL-like queries on unstructured data. The net advantage is a reduced time to market on analyzing data and a lower technical threshold on analysts to start answering business questions with the data.

Chart 19: Big Data Vendor Map



Source: First Derivatives

RapidMiner and **Predixion** are two companies providing cloud services to visualize and process data for predictive analytics. Both are desktop tools more directly integrated into modern toolsets. RapidMiner has a visual studio and tools to enable business analysts to define workflows, process and visualize data without coding. In addition, RapidMiner enables businesses to mine multiple data sources - SQL, Hadoop or Mongo from on-site sources - or provides processing and cloud storage for data, depending on what is licensed. They also have an ecosystem of partners and a marketplace for different adaptors and APIs. RapidMiner identifies sentiment analysis and risk analysis amongst its solution use cases. Predixion offers similar services to RapidMiner, but is more focused on machine learning and as a front end, utilizes Excel or other visualization tools. Predixion can work in the cloud or via an on-premises license and with multiple database types.

The broad big data technology landscape is illustrated in Chart 19.

Emerging Database Vendors – Hadoop and MongoDB

Several database pure-plays have emerged to commercialize Hadoop, with each firm having a slightly different take on the market. Most of these vendors enable technologists to download open source versions of their software and experiment on commodity servers at a very low cost. The barrier to entry for piloting a big data program hinges mainly on internal or contract expertise to configure, program and query data sets residing on these unstructured platforms:

- **Cloudera** - Founded in 2008, Cloudera was the first company to commercialize Hadoop. They package Hadoop for the enterprise and offer a proprietary enterprise management stack that includes system management, security and data governance. Their developers contribute to the open source code base, provide consulting services and train custom developers in Hadoop. Cloudera has also established an ecosystem of partners to connect and distribute their software products.

- **HortonWorks** - Founded in 2011 by a core team of engineers from Yahoo, HortonWorks was the first Hadoop company to go public on December 11, 2014. It provides similar Hadoop for enterprise software and support. Unlike Cloudera, HortonWorks' enterprise Horton Data Platform (HDP) is 100% open source and all data governance, security and management components are part of the Apache Software Foundation. HortonWorks focuses on software development and integrating Hadoop to a broader ecosystem of operating systems, infrastructure, cloud providers, applications and visualization tools. They also focus on support, relying on system integrators for professional services. HortonWorks has been a major contributor and sponsor of Apache stack development.

- **MapR** – MapR is another Hadoop-based vendor that has taken a different approach to commercializing Hadoop. They are re-architecting some of the Hadoop HDFS core to make it more suitable for enterprise use and are making it scalable to handle operational and real-time workloads. MapR is focused on optimizing resources within an enterprise data center when it comes to hosting Hadoop databases. While participating in the Hadoop open source community and retaining open source APIs, MapR believes its approach is more effective for enterprise architecture, security, managing a product road map and integrating with enterprise applications and data ecosystems. Some of this is being challenged by the growth and maturity of Hadoop in recent years.

MongoDB – Document Database

Other big data technologies have emerged and have become widely used, the most pervasive being MongoDB:

MongoDB is both a database and a company. Founded in 2007 as 10Gen, MongoDB became more of a software and open source support model company in 2009 and changed its name to MongoDB.

MongoDB is a document-oriented database or is described as a No-SQL database. Founded more as a database to operate large websites, Mongo is one of the back-ends for large web businesses, such as eBay, Foursquare and the NYTimes. The focus of Mongo is more operational and it is designed to be distributed, easy to program and perform read-write functions in a scalable, redundant way.

It uses JavaScript Object Notation 'JSON-like' documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster. MongoDB retains some of the characteristics of RMDBs, making it easier to query and port relational database data to MongoDB.

Documents can be indexed and queried and secondary indexing is also supported. MongoDB works in conjunction with Hadoop on more analytic, batch-focused use cases. Algorithmic trading, where high-speed and high throughput are required, is one use case for which MongoDB is not well suited. Such a high throughput data processing use case is more the domain of established in-memory, column databases like kdb+ (which will be discussed later). The following are use cases more suitable to MongoDB vs. Hadoop:

Mongo DB Use Cases	Hadoop Use Cases
Product/asset catalogs	Risk modeling
Geolocation	Predictive analytics
Real-time analytics	Trade surveillance
Mobile apps	Data warehouse
Social media	Ad targeting
	Churn analysis

Column-based Vendors – KX (kdb+)

Any discussion of big data within financial services and particularly investment management front offices should include mention of KX Systems (kdb+). While kdb+ has been used since 2001 to address large volume and high-velocity time-series data sets such as price, volume and position, it still needs to be recognized as a key front office big data vendor.

Like its No-SQL equivalents, kdb+ can be managed in clusters and in the cloud and is well tested in streaming data use cases, where speed and processing power are essential. Kdb+ has been used heavily in black box trading where historical data informs real-time data, and analytics, investment signals and risks are determined at high speed. It also can be connected to visualization tools to provide batch-oriented and query-based analysis. One of the advantages of this database is its hybrid ability to conduct in-memory as well as stored data processing and the fact that its application language is tightly coupled to its stream processing capabilities.

Additional big data use cases are emerging in the area of the Internet of Things (IoT), where smart meters in utilities, medical devices and oil equipment can feed kdb+ vast amounts of streaming data to be processed. Kdb+ has added additional features around geo-spatial tagging. All of these capabilities could make kdb+ a key player where sensory data can be ingested in parallel with financial data to inform investment and risk management decisions.

While still early, efforts to combine time-series data with streaming unstructured data may emerge as a powerful tool for front-office analysis of structured and unstructured data sets.

Big Data Solution Specialists

Founded in 2003, Palantir has scaled deep data science consulting with technology (Hadoop and other applications) to provide governments, healthcare and financial services firms with data analysis services and technology solutions.

More specifically, Palantir has gained capital markets traction by providing services to investment banks, including Citi, to improve market analysis productivity and provide trader surveillance capabilities.

The platform has enabled banks to join large structured and unstructured data sets from multiple data sources – siloed business-line product processors, reference data, external feeds, web data –to monitor markets, generate trade ideas, simulate trading strategies, develop hedges and assess counterparty and investment risk.

Less technical end-users and analysts can conduct analysis through their front-end and visualization tools, making the organization more innovative in its analysis and less dependent on IT for staging and preparing data.

New Technology Platforms for Enabling Big Data for Investment Managers

A new breed of platform technology start-up firms, like Incapture Technologies, is looking to make application and data management across SQL, columnar and unstructured databases possible by providing technology in a cloud service.

Incapture's Rapture platform creates a data abstraction layer to enable investment managers to write applications and interact with data from any number of database types – MySQL, Cassandra and MongoDB. This platform-as-a-service layer manages not only entitlements at the data level, but other common services – workflow, scheduling, event and resource management – and lets firms incrementally test and port applications without having to manage additional servers and databases.

Rapture was developed to create efficiencies across all types of data management tasks by consolidating Excel analysis tools, end-point applications and data stores into a centralized platform with strong entitlement controls and auditing. Some consulting or internal IT expertise is required to utilize a platform like Rapture, but built-in services are designed to lower the barrier to entry, experimentation and consolidation.

Hadoop as a Service

Another class of Software-as-a-Service companies is emerging to enable businesses to perform analytics in the cloud with Hadoop as a service. Amazon EMR has been mentioned, though there are other providers, such as Altiscale and Qubole.

Altiscale provides cloud-based always-on storage and can conduct batch and streaming data analysis from other data sources while integrating with existing BI and visualization tools. It supports much of the Apache stack and tools, such as Spark, Hive, Pig and Mahout. Pricing is focused on Hadoop usage rather than pricing for hosting. Altiscale focuses on actively managing Hadoop operations and managing Hadoop jobs in addition to being Infrastructure as a Service, so that data scientists do not have to configure nodes or manage Hadoop system administration.

Quoble offers "Big Data as a Service" if a customer already has data residing in the cloud on Amazon AWS, Google Compute Engine or Microsoft Azure services. It optimizes performance for interrogating data stores in the cloud by automating scaling of clusters, providing pre-built data integrations, scheduling and providing access to reduced cost spot pricing on Amazon's EC2 (Elastic Compute Cloud) resources. Quoble also provides Apache stack tools – Hive, HBase, Pig, Spark and Presto – as part of its managed service. So far, its cloud-based offering has penetrated the AdTech, gaming and e-commerce markets.

Offerings from Established Enterprise (Hybrid) Software and Service Providers

This survey of big data technology would not be complete without mentioning some of the larger hybrid technology and services companies that have developed big data offerings around Hadoop or other unstructured data technologies.

IBM, SAP, SAS and Microsoft to name a short list have some form of big data product and service offering and are marketing big data heavily to attract corporate business and IT mindshare in larger enterprises. They offer services at scale and are priced for corporate IT budgets, though they may not have the specialized business expertise to address the unique and more niche needs of buy-side use cases and front office research.

These companies have widely-accepted and high-standard tool sets, but their licensing structures are complex, geared towards statistics experts and priced for enterprise IT budgets. While Microsoft has more enterprise-focused offerings, such as its Microsoft Analytics Platform System on-site hosted appliances, it also has Microsoft Azure, which offers Analytics services that may be priced more in line with hedge fund IT and operational budgets.

The following table describes some of these product and service offerings:

COMPANY	PRODUCT/SERVICE	DESCRIPTION
IBM	BigInsights Suite	IBM offers various product options with its Hadoop-based BigInsights suite. They include hosted and packaged solutions, SQL on Hadoop (BigSQL), R-based tool sets, visualization and analytics tools and enterprise management features – security, cluster management and non-Hadoop application integration.
	Watson	Watson is a natural language processing, hypothesis generation and evaluation and dynamic learning system. It runs on IBM's DeepQA software which is a combination of Java, C++, Prolog and Hadoop. It is famous for being used to beat past champions in Jeopardy. While currently being used in healthcare and some financial service applications, IBM has not identified front office asset management use cases in its marketing materials.
SAP	HANA	HANA is an in-memory, column-oriented, relational database management system developed to handle both high transaction rates and complex query processing on the same platform. It is used to capture operational data in memory and analyze the data instantly. SAP offers it in cloud form as HANA Enterprise cloud service.
	SAP Sybase IQ	SAP's Sybase offering is a relational database that can scale with distributed clusters and can act as a data warehouse that co-exists with Hadoop enabling it to be structured for and loaded into Sybase for additional analysis.
	SAP Sybase Event Stream Processor (SESP)	Used in trading systems, SAP's complex event processor (CEP) can stream large data sets for real-time decision making applications.
	SAP Analytic Applications	This product suite combines HANA and SAP BusinessObjects BI suite for analysis and visualization.
	SAP InfiniteInsights	Incorporated into the SAP fold through its 2013 acquisition of Kxen, this suite of tools provides predictive analytic applications for the automated creation of predictive models. Billed as a tool for non-scientists, it lets users manage and build models quickly and will be more integrated into HANA over time.
SAS	Multiple Products adapted for Hadoop	SAS has a complex and deep array of data analysis products and has adapted their line of products for Hadoop usage, from data management to model lifecycle management to visualization. They have tools for text mining and statistical analysis. Their products are geared towards statisticians and technical business users.
Microsoft	Microsoft Analytics Platform System	Offered as an on-site appliance solution, APS includes massively parallel processing applications with their SQL Server Parallel Data Warehouse (PDW) and HDInsight – MS' Hadoop-based solution.
	Microsoft Azure - HDInsight - Machine Learning - Stream Analytics - Data Factory - Event Hubs	MS has a hosted suite of products for machine learning, stream analytics, data processing and high-volume event processing and is priced more in line with hedge fund budgets in a pay-as-you-consume model.

Disclaimer

This communication is provided by a member of the Business Advisory Services Group of Citigroup Global Markets Inc. (together with its affiliates, "Citi"). For important disclosures and disclaimers, please see https://icg.citi.com/icg/data/documents/ST_ExternalDiscl.pdf. This message is for the internal use of the intended recipients and may contain information proprietary to Citi which may not be reproduced, redistributed, or copied in whole or in part without Citi's prior consent.

The information contained in this communication is for discussion purposes only. Information provided does not constitute or include professional, legal and tax or any other form of advice and should not be relied on as such.

Information is provided to the recipient solely on the basis that the recipient will make all decisions, regardless of their nature, based on its own independent evaluation and judgment regarding their appropriateness for the recipient's own business. Any decisions made by the recipient will be made independently and separate from this communication and any other material provided by Citi, and in reliance on the advice of its other professional advisors as the recipient may deem necessary and not in reliance on any communication whether written or oral from Citi. Though Citi hopes its services will be helpful, Citi is not acting as investment advisor or fiduciary to the recipient or its clients, and the recipient's clients are not third-party beneficiaries of Citi's services. No communication whether written or oral will be understood to be an assurance or guarantee of results.

This communication is provided by Citi on a confidential basis for the recipient's use and may not be publicly disclosed. The information contained herein (a) is for informational purposes only and may not be publicly disclosed, (b) is not an offer to buy or sell any securities or service, and (c) may contain estimates and projections which may be incomplete or condensed and may be inaccurate. No representation or warranty, express or implied, is made as to the accuracy or completeness of the information and nothing herein is, or shall be relied upon as, a representation. Citi has no obligation to update or otherwise revise any such information.

IRS Circular 230 Disclosure: Citigroup Inc. and its employees are not in the business of providing, and do not provide, tax or legal advice. Any discussion of tax matters in these materials is not intended or written to be used, and cannot be used or relied upon, by any taxpayer for the purpose of avoiding tax penalties. Any such taxpayer should seek advice based on the taxpayer's particular circumstances from an independent tax advisor.

Notes