Webinar Series

Managing the Data Supply Chain

Ganesh Mani and Mehrzad Mahdavi

sup·ply chain

/səˈplī CHān/

noun

the sequence of processes involved in the production and distribution of a commodity.

Motivated by the above definition and noting that data is not necessarily a commodity (even though the comparison to oil has been made by many, as data is thought of as fueling the modern information economy), we can offer up the following broad summary:

Data Supply Chain Management is the selection, collection, organization, streamlining and flow-control of data – including any pre-processing, repair and normalization steps - to make it usable, guided by domain knowledge, for the subsequent downstream process. Typically, the next step involves analysis via traditional statistical or contemporary machine learning tools. The end goal of the exercise is to generate insights that can imply customer value, inform revenue or pricing metrics, optimize costs and help gain a competitive advantage in the marketplace.

Why is the Data Supply Chain important? The outcome of analytical engines in general and machine learning in particular can be highly dependent on the quality and other attributes of the ingested data. The exponential increase in the quantity and variety of data provides opportunities as well as challenges with regards to the sourcing, selection, and preparation (cleansing, aggregating, and normalizing) of the data. Another dimension of the data supply chain is its integrity as well as compliance with many evolving regulations regarding privacy, security and ethical use. Trustworthy data is key to trusted outcomes.

The moniker 'Big Data' has been popularized over the last decade. In the following section, we describe its nuances, especially as it relates to performing due diligence on the data supply chain.

Key data dimensions:

These dimensions are considered standard across many domains.

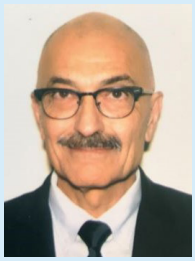
- ***Volume:*** The amount of data collected and stored via records, transactions, tables, files, etc.
- ***Velocity:*** The speed at which data is sent or received. Data can be streamed; or, received in batch mode, real-time or near-real-time.
- ***Variety:*** Data often comes from a variety of sources and in various formats – be it structured (such as SQL tables or CSV / Excel numeric files), semi-structured (such as documents in XML, HTML or JSON format), or unstructured (such as a blog post or video).

Big data, especially the recently-dubbed alternative variety, in investment management can be further classified based on how it was generated: *produced by individuals* (such as social media posts), generated through *business processes* (such as e-commerce or credit card transactions data), or *generated by sensors* (e.g., via radar or satellite imagery, sensors installed in industrial facilities or on equipment, drones). Datasets generated by individuals are often in the unstructured textual format and commonly require natural language processing. Sensor-generated data can be produced in both structured and unstructured formats. Many business-generated datasets, such as credit card transactions, and company 'exhaust' data may need to comport with existing and emerging legal and privacy considerations.

In addition to the source of data, collection techniques can be passive or active, including proactively seeking additional elements. Additional data elements may be collected, following a cost-benefit analysis – e.g., when it is estimated that spending $X on additional data collection and processing may result in a benefit exceeding $2X!

Other attributes of data (besides the preceding classification) may be important from an intended use standpoint. Investment professionals may want to map a dataset to an asset class or investment style, after considering its quality, technical specifications and alpha potential. The following related issues come to mind and will be discussed:

- ***Asset classes.*** Most ready-available datasets are still focused on equities and commodities. There is relatively little alternative data on interest rates and currencies, making such data sets more valuable to select investors.
- ***Investment styles.*** Much of the data is sector- and stock-specific, germane to equity long-short investors. There is also a significant amount of data relevant for macro investors (such as consumer credit, China economic activity and shipping volumes). Certain alternative datasets can be used to substitute traditional metrics of market risk, and



FDP Leadership Team
Dr. Mahdavi, Ph.D.,
Executive Director
FDP Institute

Dr. Mahdavi is a technology entrepreneur with focus on breakthrough digital transformation for the Fintech and the energy sectors. He is a recognized expert and frequent keynote speaker on application of AI, IoT, and Cloud computing in financial sector and industries.

Dr. Mahdavi managed major global businesses in the energy sector. He is currently the executive director of the Financial Data Professionals Institute (FDPI), a non-profit organization founded with CAIA. Mehrzad holds a PhD in Nuclear Science and Technology from the University of Michigan a Bachelor of Science in Electrical and Electronics Engineering from the University of Illinois at Urbana-Champaign.

FDP Advisory Board
Ganesh Mani, Ph.D.
Adjunct Faculty
Carnegie Mellon

Ganesh Mani is on the adjunct Faculty of Carnegie Mellon University and is considered a thought leader in the areas of investment management and AI / FinTech. He has been a pioneer in applying innovative techniques to multiple asset classes and has worked with large asset management firms incl. hedge funds, after having sold one of the earliest AI/ML-based investment management boutiques into SSgA (in the late nineties), nucleating the Advanced Research Center there.

Ganesh has been featured on ABC Nightline and in a Barron's cover story titled "New Brains on the Block". Mr. Mani has an MBA in Finance and a PhD in Artificial Intelligence from the University of Wisconsin-Madison, as well as an undergraduate degree in Computer Science from the Indian Institute of Technology, Bombay. Ganesh is a charter member of TiE (www.TiE.org), was an early member of the Chicago Quantitative alliance and is on the advisory board of the Journal of Financial Data Science.

Managing the Data Supply Chain

Ganesh Mani and Mehrzad Mahdavi

some signals are relevant only for high-frequency quant traders. Perhaps the most important data attribute is its uniqueness and hence potential.

- ***Alpha content***. Alpha content has to be analyzed relative to the purchase or trade entry price and implementation cost of deploying the dataset.
- ***How well-known* is a dataset?** The more popular a dataset is, the less likely it is to lead to a stand-alone strategy with a robust Sharpe ratio. Well-known public datasets consisting of plain-vanilla financial ratios (P/E, P/B, etc.) likely have fairly low alpha content and are not viable as standalone strategies (They may still be useful in a diversified risk-premia portfolio.). New datasets – for instance, capturing news and social media sentiment relating to a company's new products or customer service - are starting to emerge, but their utility needs to be evaluated in the context of other data, that it is used alongside.
- ***Stage of processing of data* when acquired.** Fundamental investors prefer processed signals and insights rather than a large amount of raw, hard-to-wrangle data. Most big/alternative datasets come in a semi-processed format. However, its alpha content and mapping to tradable instruments need to be assessed; in addition, it may be necessary to handle outliers and seasonality effects. Finally, raw data – devoid of domain annotations - is likely to be of little use to most investors.
- ***Quality of data,* especially for data scientists and quants.** Missing data or outliers are an important consideration. If data has been backfilled, the method of imputing missing values must be clarified. Knowing if the missing data holes were random or followed some patterns is also useful in the model-building stage.
- ***Frequency and latency.*** Frequency of data can be, for instance, intra-day, daily, weekly, or quarterly. The latency of data can be due to collection, operational or legal constraints.
- ***Provenance***. Can be qualified via source authority, location / geography and by demographics.

Classification and attribute tags are part of a comprehensive taxonomy framework that needs to be established to label target data and choose among categories of machine learning algorithms (to initially deploy on the data to elicit insights and / or imply alpha).

Why the buzz around Alternative Data in particular and Big Data in general? Alternative data promises to provide an "edge" for the investment professionals. As more investors adopt alternative datasets, the market will start reacting faster and will increasingly anticipate traditional or 'old' data sources such as quarterly corporate earnings or low-frequency macroeconomic data. This change gives a potential edge to quant managers and those willing to adapt and learn about new datasets and methods, deploying them rapidly. Eventually, 'old' datasets will lose most of their predictive value, and new alternative datasets - that anticipate official, reported values - may increasingly become standardized. There will be an ongoing effort to uncover new higher-frequency datasets and refine/supplement old ones. Machine learning techniques will become a standard tool for quantitative investors and perhaps some fundamental investors too. Systematic strategies deployed by quants such as *risk premia, trend following,* and *equity long-short* will increasingly adopt machine learning tools and methods, acting on information, a majority of which may have its origin in alternative datasets.

Hedge fund managers are positioned to get much value from big data. While big data may provide the most value for short-term investors, long-term institutional investors can also benefit from its systematic use. More importantly, they may not have to compete with hedge fund managers to acquire alternative data sets that are most suited for short-term trading. By focusing on defensive or longer-duration alpha strategies, long-term investors should be able to acquire alternative datasets at reasonable costs, as they may not necessarily be in high demand by the trading-oriented asset managers. For example, ESG investing is becoming a major force, with the twin goals of medium-term returns and long-term positive impact on the planet and society.

ESG-aware investors may consider a) **environmental** factors such as energy use and efficiency, animal welfare and deforestation implications b) **societal** factors such as human rights and working conditions, diversity of workforce and sentiment in the local community; and c) **governance** factors such as board structure and composition; minority shareholder rights, executive compensation and tax strategy. Markers in the data supply chain – in addition to the specific data collected for these purposes – may be able to imply a nuanced perspective with respect to these factors.

Big data and machine learning algorithms could also be used to improve compliance at the firm level and to aid in performing due diligence on managers hired by long-term investors.

As pointed out by Lopez de Prado in a Bloomberg article[1], the majority of the hedge funds that have consistently beaten the market averages in recent years have employed highly mathematical, data-intensive strategies. In the same article, it is reported that several traditional hedge funds have closed, while quantitative funds such as Renaissance Technologies[2] and Two Sigma are still attracting new assets and clients.

In summary, the data supply chain is something a traditional as well as alternative portfolio analyst / manager should pay attention to. It can help inform security selection, asset allocation, portfolio tilting (towards desired factors) and risk management tasks, to name a few.

[1] https://www.bloomberg.com/news/articles/2018-10-09/the-big-problem-with-machine-learning-algorithms

[2] *The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution* by Gregory Zuckerman

Additional companion reference:

Alternative Data: Don't bark up the wrong tree (too early)!

By Ganesh Mani, PhD

The Financial Data Professional Institute (FDPI) was established by the CAIA Association to address the growing need in finance for a workforce that has the skills to perform in a digitized world where an increasing number of decisions will be data and analytics driven. The FDP curriculum introduces candidates to central concepts of machine learning and big data, including ethical and privacy issues, and their roles in various segments of the financial industry to boost and integrate quant knowledge into analytics' skills.