The Book of Alternative Data

A Guide for Investors, Traders, and Risk Managers

Alexander Denev
Saeed Amen

WILEY

"Alternative data is one of the hottest topics in the investment management industry today. Whether it is used to forecast global economic growth in real time, to parse the entrails of a company with more granularity than that offered by a quarterly report, or to better understand stock market behaviour, alternative data is something that everyone in asset management needs to get to grips with. Alexander Denev and Saeed Amen are able guides to a convoluted subject with many pitfalls, both technical and theoretical, even for those who still think Python is a snake best avoided."

— Robin Wigglesworth, Global finance correspondent, Financial Times.

"Congratulations to the authors for producing such a timely, comprehensive, and accessible discussion of alternative data. As we move further into the twenty-first century, this book will rapidly become the go-to work on the subject."

— Professor David Hand, Imperial College London

"Over the last decade, alternative data has become central to the quest for temporary monopoly of information. Yet, despite its frequent use, little has been written about the end-to-end pipeline necessary to extract value. This book fills the omission, providing not just practical overviews of machine learning methods and data sources, but placing as much importance on data ingestion, preparation, and pre-processing as on the models that map to outcomes. The authors do not consider methodology alone, but also provide insightful case studies and practical examples, and highlight the importance of cost-benefit analysis throughout. For value extraction from alternative data, they provide informed insights and deep conceptual understanding – crucial if we are to successfully embed such technology at the heart of trading."

— Stephen Roberts, Royal Academy of Engineering/Man Group Professor of Machine Learning, University of Oxford, UK, and Director of the Oxford-Man Institute of Quantitative Finance

"True investment outperformance comes from the triad of data plus machine learning plus supercomputing. Alexander Denev and Saeed Amen have written the first comprehensive exposition of alternative data, revealing sources of alpha that are not tapped by structured datasets. Asset managers unfamiliar with the contents of this book are not earning the fees they charge to investors."

— Dr. Marcos López de Prado, Professor of Practice at Cornell University, and CIO at True Positive Technologies LP

"Alexander and Saeed have written an important book about an important topic. I am involved with alternative data every day, but I still enjoyed the perspectives in the book, and learned a lot. I highly recommend it to everybody looking to harness the power of alt data (and avoid the pitfalls!)."

— Jens Nordvig, Founder and CEO of Exante Data

The Book of Alternative Data

A Guide for Investors, Traders, and Risk Managers

ALEXANDER DENEV
SAEED AMEN

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:
Names: Denev, Alexander, author. | Amen, Saeed, 1982- author.
Title: The book of alternative data : a guide for investors, traders and risk managers / Alexander Denev, Saeed Amen.
Description: Hoboken, New Jersey : Wiley, [2020] | Includes bibliographical references and index.
Identifiers: LCCN 2020008783 (print) | LCCN 2020008784 (ebook) | ISBN 9781119601791 (hardback) | ISBN 9781119601814 (adobe pdf) | ISBN 9781119601807 (epub)
Subjects: LCSH: Investments | Financial risk management. | Big data.
Classification: LCC HG4529 .D47 2020 (print) | LCC HG4529 (ebook) | DDC 332.63/204—dc23
LC record available at https://lccn.loc.gov/2020008783
LC ebook record available at https://lccn.loc.gov/2020008784

Cover Design: Wiley
Cover Image: © akindo/Getty Images

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Natalie, with all my love. –Alexander
For Gido and Baba, in life, in time, in spirit, your path is forever my guide. –Saeed

Contents

Preface

Data permeates through our world, in ever increasing amounts. This fact alone is not sufficient for data to be useful. Indeed, data has no utility, if it is devoid of information, which could aide our understanding. Data needs to be insightful for it to be of use and it also needs to be processed in the appropriate way. In the pre-Big Data age days, statistics such as averages, standard deviation, correlations were calculated on structured datasets to illuminate our understanding of the world. Models were calibrated on (a small number of) input variables which were often well "understood" to obtain an output via well-trodden methods like, say, linear regression.

However, interpreting Big Data, and hence alternative data, comes with many challenges. Big Data is characterized by properties such as volume, velocity and variety and other Vs, which we will discuss in this book. It is impossible to calculate statistics, unless datasets are well structured and relevant features are extracted. When it comes to prediction, the input variables derived from Big Data are numerous and traditional statistical methods can be prone to overfitting. Moreover, nowadays calculating statistics or building models on this data must be done sometimes frequently and in a dynamic way to account for the always changing nature of the data in our high frequency world.

Thanks to technological and methodological advances, understanding Big Data and by extension alternative data, has become a tractable problem. Extracting features from messy enormous volumes of data is now possible thanks to the recent developments in artificial intelligence and machine learning. Cloud infrastructure allows elastic and powerful computation to manage such data flows and to train models both quickly and efficiently. Most of the programming languages in use today are open source and many such as Python have a large number of libraries in the sphere of machine learning and data science more broadly, making it easier to develop tech stacks to number crunch large datasets.

When we decided to write this book, we felt that there was a gap in the book market in this area. This gap seemed at odds with the ever growing importance of data, and in particular, alternative data. We live in a world, which is rich with data, where many datasets are accessible and available at a relatively low cost. Hence, we thought that it was worth writing a lengthy book to address how to address the challenges of

how to use data profitably. We do admit though that the world of alternative data and its use cases is and will be subject to change in the near future. As a result, the path we paved with this book is also subject to change. Not least the label "alternative data" might become obsolete as it could soon turn mainstream. Alternative data may simply become "data". What might seem to be great technological and methodological feats today to make alternative data usable, may soon become trivial exercises. New datasets from sources we could not even imagine could begin to appear, and quantum computing could revolutionise the way we look at data.

We decided to target this book at the investment community. Applications, of course, can be found elsewhere, and indeed everywhere. By staying within the financial domain, we could also have discussed areas such as credit decisions or insurance pricing, for example. We will not discuss these particular applications in this book, as we decided to focus on questions that an investor might face. Of course, we might consider adding these applications in future editions of the book.

At the time of writing, we are living in a world afflicted by COVID-19. It is a world, in which it is very important for decision makers to make the right judgement, and furthermore, these decisions must be done in a timely manner. Delays or poor decision making can have fatal consequences in the current environment. Having access to data streams that track the foot traffic of people can be crucial to curb the spread of the disease. Using satellite or aerial images could be helpful to identify mass gatherings and to disperse them for reasons of public safety. From an asset manager's point of view, creating nowcasts before official macroeconomic figures and company financial statements are released, results better investment decisions. It is no longer sufficient to wait several months to find out about the state of the economy. Investors want to have be able to estimate such points on a very high frequency basis. The recent advances in technology and artificial intelligence makes all this possible.

So, let us commence on our journey through alternative data. We hope you will enjoy this book!

Acknowledgments

We would like to thank our friends and colleagues who have helped us by providing suggestions and correcting our errors.

In first place, we would like to express our gratitude to Dr. Marcos Lopez de Prado who gave us the idea of writing this book. We would like to thank Kate Lavrinenko without whom the chapter on outliers would not have been possible; Dave Peterson, who proofread the entire book and provided useful and thorough feedback; Henry Sorsky for his work with us on the automotive fundamental data and missing data chapters, as well as proofreading many of the chapters and pointing out mistakes; Doug Dannemiller for his work around the risks of alternative data which we leveraged; Mike Taylor for his contribution to the data vendors section; Jorge Prado for his ideas around the auctions of data.

We would also like to extend our thanks to Paul Bilokon and Matthew Dixon for their support during the writing process. We are very grateful to Wiley, and Bill Falloon in particular, for the enthusiasm with which they have accepted our proposal, and for the rigor and constructive nature of the reviewing process by Amy Handy. Last but not least, we are thankful to our families. Without their continuous support this work would have been impossible.

The Book of Alternative Data

PART 1

Introduction and Theory

CHAPTER 1

Alternative Data: The Lay of the Land

1.1 INTRODUCTION

There is a considerable amount of buzz around the topic of alternative data in finance. In this book, we seek to discuss the topic in detail, showing how alternative data can be used to enhance understanding of financial markets, improve returns, and manage risk better.

This book is aimed at investors who are in search of superior returns through nontraditional approaches. These methods are different from fundamental analysis or quantitative methods that rely solely on data widely available in financial markets. It is also aimed at risk managers who want to identify early signals of events that could have a negative impact, using information that is not present yet in any standard and broadly used datasets.[1]

At the moment of writing there are mixed opinions in the industry about whether alternative data can add any value in the investment process on top of the more standardized data sources. There is news in the press about hedge funds and banks who have tried, but failed to extract value from it (see e.g. Risk, 2019). We must stress, however, that the absence of predictive signals in alternative data is only one of the components of a potential failure. In fact, we will try to convince the reader, through the practical examples that we will examine, that useful signals can be gleaned from alternative data in many cases. At the same time, we will also explain why any strategy that aims to extract and make successful use of signals is a combination of algorithms, processes, technology, and careful cost-benefit analysis. Failure to tackle any of these aspects in the right way will lead to a failure to extract usable insights from alternative data. Hence, the proof of the existence of a signal in a dataset is not sufficient

[1] A lot of applications of alternative data are being found today in insurance and credit markets (see e.g. Turner, 2008; Turner, 2011; Financial Times, 2017). We will not explicitly treat them here, although the alternative data generalities we will examine are also applicable to those cases.

to benefit from a superior investment strategy, given that there are many other subtle issues at play, most of which are dynamic in nature, as we will explain later.

In this book, we will also discuss in detail the techniques that can be used to make alternative data usable for the purposes we have already noted. These will be techniques belonging to what are labeled today as the fields of Machine Learning (ML) and Artificial Intelligence (AI). However, we do not want to give the upfront impression of being unnecessarily complex, with these "sophisticated" catchall terms. Hence, we will also include simpler and more traditional techniques, such as linear and logistic regression,[2] with which the financial community is already familiar. Indeed, in many instances simpler techniques can be very useful when seeking to extract signals from alternative datasets in finance. Nevertheless, this is not a machine learning textbook and hence we will not delve in the details of each technique we will use, but we will only provide a succinct introduction. We will refer the reader to the appropriate texts where necessary.

This is also not a book about the technology and the infrastructure that underlie any real-world implementations of alternative data. These topics encompassing data engineering are still, of course, very important. Indeed, they are necessary for anything found to be a signal in the data to be of any use in real life. However, given the variety and the deep expertise needed to treat them in detail, we believe that these topics deserve a book on their own. Nevertheless, we must stress that methodologies that we use in practice to extract a signal are often constrained by technological limitations. Do we need an algorithm to work fast and deliver results in almost real time or can we live with some latency? Hence, the type of algorithm we choose will be very much determined by technological constraints like these. We will hint at these important aspects throughout, although this book will not be, strictly speaking, technological.

In this book, we will go through practical case studies showing how different alternative data sources can be profitably employed for different purposes within finance. These case studies will cover a variety of data sources and for each of them will explore in detail how to solve a specific problem like, for example, predicting equity returns from fundamental industrial data or forecasting economic variables from survey indices. The case studies will be self-contained and representative of a wide array of situations that could appear in the real-world applications, across a number of different asset classes.

Finally, this book will not be a catalogue of all the alternative data sources existing at the moment of writing. We deem this to be futile because, in our dynamic world, the number and variety of such datasets increase every day. What is more important, in our view, is the process and techniques of how to make the available data useful. In doing so, we will be quite practical by also examining mundane problems that appear in sieving through datasets, the missteps and mistakes that any practical application entails.

This book is structured as follows. Part I will be a general introduction to alternative data, the processes and the techniques to make it usable in an investment strategy. In Chapter 1, we will define alternative data and create a taxonomy. In Chapter 2

[2] In fact, most of the ML/AI textbooks start with these simple techniques.

we will discuss the subtle problem of how to price datasets. This subject is currently being actively debated in the industry. Chapter 3 will talk about the risks associated with alternative data, in particular the legal risks, and we will also delve more into the details of the technical problems that one faces when implementing alternative data strategies. Chapter 4 introduces many of the machine learning and structuring techniques that can be relevant for understanding alternative data. Again, we will refer the reader to the appropriate literature for a more in-depth understanding of those techniques.

Chapter 5 will examine the processes behind the testing and the implementation of alternative data signals-based strategies. We will recommend a fail-fast approach to the problem. In a world where datasets are many and further proliferating, we believe that this is the best way to proceed.

Part II will focus on some real-world use cases, beginning with an explanation of factor investing in Chapter 6, and a discussion of how alternative data can be incorporated in this framework. One of the use cases will not be directly related to an investment strategy but is a problem at the entry point of any project and must be treated before anything else is attempted – missing data, in Chapters 7 and 8. We also address another ubiquitous problem of outliers in data (see Chapter 9). We will then examine use cases for investment strategies and economic forecasting based on a broad array of different types of alternative datasets, in many different asset classes, including public markets such as equities and FX. We also look at the applicability of alternative data to understand private markets (see Chapter 20), where markets are typically opaquer given the lack of publicly available information. The alternative datasets we shall discuss include automotive supply chain data (see Chapter 10), satellite imagery (see Chapter 13), and machine readable news (see Chapter 15). In many instances, we shall also illustrate the use case with trading strategies on various asset classes.

So, to start this journey, let's explain a little bit more about what the financial community means by "alternative data" and why it is considered to be such a hot topic.

1.2 WHAT IS "ALTERNATIVE DATA"?

It is widely known that information can provide an edge. Hence, financial practitioners have historically tried to gather as much data as is feasible. The nature of this information, however, has changed over time, especially since the beginning of the Big Data revolution.[3] From "standard" sources like market prices and balance sheet information, it evolved to include others, in particular those that are not strictly speaking financial. These include, for example, satellite imagery, social media, ship

[3]There is no precise date of when this revolution started, and certainly this has not been an instantaneous event. In *Thank You for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations,* Thomas Friedman puts the starting year as 2007 because this is the year when major development in computational power, software, sensors, and connectivity happened. The term "Big Data" has been around since the 1990s and the father of the term is John Mashey, who was the chief scientist at Silicon Graphics at the time.

movements, and the Internet-of-Things (IoT). The data from these "nonstandard" sources is labeled alternative data.

In practice, alternative data has several characteristics, which we list below. It is data that has at least one of the following features:

- Less commonly used by market participants
- Tends to be more costly to collect, and hence more expensive to purchase
- Usually outside of financial markets
- Has shorter history
- More challenging to use

We must note from this list that what constitutes alternative data can vary significantly over time according to how widely available it is, as well has how embedded in a process it is. Obviously, today most financial market data is far more commoditized and more widely available than it was decades ago. Hence, it is not generally labeled as alternative. For example, a daily time series for equity closing prices is easily accessible from many sources and it is considered nonalternative. In contrast, very high frequency FX data, although financial, is far more expensive, specialized, and niche. The same is also true of comprehensive FX volume and flow data, which is less readily available. Hence, these market derived datasets may then be considered alternative. The cost and availability of a dataset are very much dependent on several factors, such as asset class and frequency. Hence, these factors determine whether the label "alternative" should be attached to it or not. Of course, clear-cut definitions are not possible and the line between "alternative" and "nonalternative" is somewhat blurred. It is also possible that, in the near future, what we consider "alternative" will become more standardized and mainstream. Hence, it could lose the label "alternative" and simply be referred to as data.

In recent years, the alternative data landscape has significantly expanded. One major reason is that there has been a proliferation of devices and processes that generate data. Furthermore, much of this data can be recorded automatically, as opposed to requiring manual processes to do so. The cost of data storage is also coming down, making it more feasible to record this data to disk for longer periods of time. The world is also awash with "exhaust data," which is data generated by processes whose primary purpose is not to collect or generate and sell the data. In this sense, data is a "side effect." The most obvious example of exhaust data in financial markets is market data. Traders trade with one another on an exchange and on an over-the-counter basis. Every time they post quotes or agree to trade at a price with a counterparty, they create a data point. This data exists as an exhaust of the trading activity. The concept of distributing market data is hardly new and has been an important part of markets for the ages and is an important part of the revenue for exchanges and trading venues.

However, there are other types of exhaust data that have been less commonly utilized. Take, for example, a large newswire organization. Journalists continually write news articles to inform their readers as part of their everyday business.

This generates large amounts of text daily, which can be stored on disk and structured. If we think about firms such as Google, Facebook, and Twitter, their users essentially generate vast amounts of data, in terms of their searches, their posts, and likes. This exhaust data, which is a by-product of user activity, is monetized by serving advertisements targets toward users. Additionally, each of us creates exhaust data every time we use our mobile phones, creating a record of our location and leaving a digital footprint on the web.

Corporations that produce and record this exhaust data are increasingly beginning to think about ways of monetizing it outside of their organization. Most of the exhaust data, however, remains underutilized and not monetized. Laney (2017) labels this "dark data." It is internal, usually archived, not generally accessible and not structured sufficiently for analysis. It could be archived emails, project communications, and so on. Once such data is structured, it will also make that data more useful for generating internal insights, as well as for external monetization.

1.3 SEGMENTATION OF ALTERNATIVE DATA

As already mentioned, we will not describe all the sources of alternative data but will try to provide a concise segmentation, which should be enough to cover most of the cases encountered in practice. First, we can divide the alternative data sources into the following high-level categories of generators:[4] individuals, institutions[5] and sensors, and derivations or combinations of these. The latter is important because it can lead to the practically infinite proliferation of datasets. For example, a series of trading signals extracted from data can be considered as another transformed dataset.

The collectors of data can be either institutions or individuals. They can store information created by other data generators. For example, credit card institutions can collect transactions from individual consumers. Concert venues could use sensors to track the number of individuals entering a particular concert hall. The data collection can be either manual or automatic (e.g. handwriting versus sensors). The latter is prevalent in the modern age, although until a couple of decades ago the opposite was true.[6] The data recorded can either be in a digital or analog form. This segmentation is summarized in Table 1.1.

We can further subdivide the high-level categories into finer-grained categories according to the type of data is generated. A list can never be exhaustive. For example, individuals generate internet traffic and activity, physical movement and location (e.g. via mobile phone), and consumer behavior (e.g. spending, selling); institutions generate reports (e.g. corporate reports, government reports), institutional

[4] Here we draw inspiration from the United Nations classification (see United Nations, 2015), although, in this text, we make the distinction between generators and collectors.

[5] By "institutions" we mean associations of individuals such as corporations, public entities, or governments.

[6] This consideration might be important if we want to enrich short time series with previous and old recordings (e.g. temperature or river levels time series going as far back as the 19th century), or loss on loans in banks in the 1990s for loss-given-default (LGD) modeling.

TABLE 1.1 Segmentation of alternative data.

Who Generates the Data?	Who Collects the Data?	How Is It Collected?	How Is It Recorded?
Physical processes	Individuals	Manually	Via digital methods
Individuals	Institutions	Automatically	Via analog methods
Institutions			

behavior (e.g. market activity); and physical processes collect information about physical variables (e.g. temperature or luminosity, which can be detected via sensors).

As individuals, we generate data via our actions: we spend, we walk, we talk, we browse the web, and so on. Each of these activities leaves a digital footprint that can be stored and later analyzed. We have limited action capital, which means that the number of actions we can perform each day is limited. Hence, the amount of data we can generate individually is also limited by this. Institutions also have limited action capital: mergers and acquisitions, corporate reports, and the like. Sensors also have limited data generation capacity given by the frequency, bandwidth, and other physical limitations underpinning their structure. However, data can also be artificially generated by computers that aggregate, interpolate, and extrapolate data from the previous data sources. They can transform and derive the data as already mentioned above. Therefore, for practical purposes we can say that the amount of data is unlimited. One such example of data generated by a computer is that of an electronic market maker, which continually trades with the market and publishes quotes, creating a digital footprint of its trading activity.

How to navigate this infinite universe of data and how to select which datasets we believe might contain something valuable for us is almost an art. Practically speaking, we are limited by time and budget constraints. Hence, venturing into inspecting many data sources, without some process of prescreening, can be risky and is also not cost effective. After all, even "free" datasets have a cost associated with them, namely the time and effort spent to analyze them. We will discuss how to approach this problem of finding datasets later and how a new profession is emerging to tackle this task – the data scout and data strategist.

Data can be collected by firms and then resold to other parties in a raw format. This means that no or minimal data preprocessing is performed. Data can be then processed by cleansing it, running it through quality control checks, and maybe enriching it through other sources. Processed data can then be transformed into signals to be consumed by investment professionals.[7] When data vendors do this processing, they can do it for multiple clients, hence reducing the cost overall.

These signals could be, for example, a factor that is predictive of the return of an asset class or a company, or an early warning indicator for an extreme event.

[7]There are potentially different degrees of the data being processed. In this sense, data can be also semi-processed. We will not use this fine distinction here, but this is something to bear in mind.