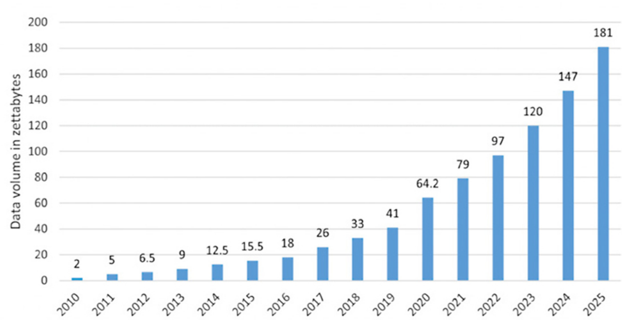
CRC PRESS/ OPTIRISK SERIES IN FINANCE

Handbook of Alternative Data in Finance

Volume I

Edited by
Gautam Mitra, Christina Erlwein-Sayer,
Kieu Thi Hoang, Diana Roman, and Zryan Sadik



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

OptiRisk
SYSTEMS

"Alternative data has become a hot topic in finance. New kinds of data, new data sources, and of course new tools for processing such data offer the possibility of new and previously unsuspected signals. In short alternative data lead to the promise of enhanced predictive power. But such advance does not come without its challenges - in terms of the quality of the data, the length of its history, reliable data capture, the development of appropriate statistical, AI, machine learning, and data mining tools, and, of course, the ethical challenges in the face of increasingly tough data protection regimes. Gautam Mitra and his colleagues have put together a superb collection of chapters discussing these topics, and more, to show how alternative data, used with care and expertise, can reveal the bigger picture."
– Professor David J. Hand, Emeritus Professor of Mathematics and Senior Research Investigator, Imperial College, London

"Digital capital is now so important that it can rightly be viewed as a factor of production, especially in the financial sector. This handbook does for the field of alternative data what vendors of alternative data do for data itself; and that is to provide structure, filter noise, and bring clarity. It is an indispensable work which every financial professional can consult, be it for an overview of the field or for specific details about alternative data."
– Professor Hersh Shefrin, Mario L. Belotti Professor of Finance, Santa Clara University

An impressive and timely contribution to the fast developing discipline of data driven decisions in the trading and management of financial risk. Automated data collection, organization, and dissemination is part and parcel of Data Science and the Handbook covers the current breadth of these activities, their risks, rewards, and costs. A welcome addition to the landscape of quantitative finance.
–Professor Dilip Madan, Professor of Finance, Robert H. Smith School of Business

"The Handbook of Alternative Data in Finance is the most comprehensive guide to alternative data I have seen. It could be called the Encyclopaedia of Alternative Data. It belongs to the desktop, not the bookshelf, of every investor."
– Ernest Chan, Respected Academic, Author, Practicing Fund Manager, Entrepreneur and Founder of PredictNow.AI

"Professor Gautam Mitra and his team unpack the topic of alternative data in finance, an ambitious endeavor given the fast-expanding nature of this new and exciting space. Alternative data powered by Natural Language Processing and Machine Learning has emerged as a new source of insights that can help investors make more informed decisions, stay ahead of competition and mitigate emerging risks. This handbook provides a strong validation of the substantial added value that alternative data brings. It also helps promote the idea that data driven decisions are better and more sustainable – something we, at RavenPack, firmly believe."
– Armando Gonzalez, CEO and Founder of RavenPack

"As the 1st Duke of Marlborough, John Churchill, wrote in 1715: 'No war can be conducted successfully without early and good intelligence.' The same can be said for successful trading. In that light, the Handbook of Alternative Data in Finance contains vital insights about how to gather and use alternative data —in short, intelligence —to facilitate successful trading."
– Professor Steve H. Hanke, Professor of Applied Economics, The Johns Hopkins University, Baltimore, USA

"*The Handbook of Alternative Data in Finance* is cutting edge and it bridges a huge gap in the representative studies on emerging areas of finance where alternative data can be profitably utilised for better informed decisions. The practical insights in the book would come very handy to both investors and researchers who look for fresh ideas."
– Ashok Banerjee, Director, Indian Institute of Management Udaipur, Formerly Dean, and Faculty-in-charge of the Finance Lab at Indian Institute of Management Calcutta

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Handbook of Alternative Data in Finance
Volume I

Handbook of Alternative Data in Finance, Volume I motivates and challenges the reader to explore and apply Alternative Data in finance. The book provides a robust and in-depth overview of Alternative Data, including its definition, characteristics, difference from conventional data, categories of Alternative Data, Alternative Data providers, and more. The book also offers a rigorous and detailed exploration of process, application and delivery that should be practically useful to researchers and practitioners alike.

Features

- Includes cutting-edge applications in machine learning, fintech, and more
- Suitable for professional quantitative analysts, and as a resource for postgraduates and researchers in financial mathematics
- Features chapters from many leading researchers and practitioners.

CRC Press/OptiRisk Series in Finance

Series Editors:
Gautam Mitra
OptiRisk Systems

Forthcoming books:

Handbook of Alternative Data in Finance, Volume I
Edited by Gautam Mitra, Christina Erlwein-Sayer, Kieu Thi Hoang, Diana Roman, and Zryan Sadik

For more information about the series, please visit: https://www.routledge.com/CRC-PressOptiRisk-Series-in-Finance/book-series/ORSF

Handbook of Alternative Data in Finance

Volume I

Edited by
Gautam Mitra
Christina Erlwein-Sayer
Kieu Thi Hoang
Diana Roman
Zryan Sadik

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Designed cover image: ©ShutterStock Images

First edition published 2023
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

Names: Mitra, Gautam, editor. | Erlwein-Sayer, Christina, editor. | Hoang, Kieu Thi, editor. | Roman, Diana, editor. | Sadik, Zryan, editor.
Title: Handbook of alternative data in finance / edited by Gautam Mitra, Christina Erlwein-Sayer, Kieu Thi Hoang, Diana Roman, Zryan Sadik.
Description: Boca Raton : C&H/CRC Press, 2023. | Series: CRC Press/OptiRisk series in finance | "A Chapman & Hall book." | Includes bibliographical references and index.
Identifiers: LCCN 2022060251 (print) | LCCN 2022060252 (ebook) | ISBN 9781032276489 (hardback) | ISBN 9781032276700 (paperback) | ISBN 9781003293644 (ebook)
Subjects: LCSH: Finance--Data processing. | Finance--Mathematical models.
Classification: LCC HG101 .H34 2023 (print) | LCC HG101 (ebook) | DDC 332.028541--dc23/eng/20230417
LC record available at https://lccn.loc.gov/2022060251
LC ebook record available at https://lccn.loc.gov/2022060252

ISBN: 978-1-032-27648-9 (hbk)
ISBN: 978-1-032-27670-0 (pbk)
ISBN: 978-1-003-29364-4 (ebk)

DOI: 10.1201/9781003293644

Typeset in Latin Modern font
by KnowledgeWorks Global Ltd.

Publisher's note: This book has been prepared from camera-ready copy provided by the authors

Contents

Preface

SETTING THE SCENE: BACKGROUND

We at OptiRisk have set out to research, acquire knowledge and be experts in the domain of News Analytics, Sentiment Analysis and Alternative Data in Finance. Our journey started with The Handbook of News Analytics in Finance (2011), which was followed by The Handbook of Sentiment Analysis in Finance (2016). We have stayed the course and as a Financial Analytics company continued to research in the domain of trading and fund management; this has culminated in our latest work: The Handbook of Alternative Data in Finance (2022). We wish to set a context for this multiauthor volume of the Handbook. In a Tome by a single author, one might get a fair depth of analysis. We refer the readers to (Denev & Amen, 2020),[1] Mahjouri (Mahjouri, 2023),[2] and Neptune Knowledge (Knowledge, 2022).[3] All these books discuss how Alternative Data is used to predict market movements make smarter investment decisions and risk control decisions. But these also invariably include the author's biases and a single perspective. In a multiauthor book, however, you get a 360-degree view and a wider perspective. In applied finance, an area of empirical research, this provides a better value to the market participants; specially, to traders, analysts, fund managers and others (Mitra et al., 2020).[4] The target audience of the Handbook is the latter group of financial market professionals as well as academics who are interested in the financial markets.

RECENT PROGRESS

Over the last 50 years, economists, operations researchers and management science and business analysts have researched extensively the topic of "models for decision making". In the early days, the role of data in such models were not even considered. Later from 1990s role of data was acknowledged but it

[1]Denev, A. & Amen, S., 2020. *The Book of Alternative Data: A Guide for Investors, Traders and Risk Managers.* s.l.:Wiley.

[2]Mahjouri, M., 2020. *Alternative Data: Capturing the Predictive Power of Big Data for Investment Success.* s.l.:Willey.

[3]Knowledge, N., 2022. *Introduction to Alternative Data: The power of data; a book for company executives and investment professionals*, s.l: Amazon

[4]Mitra, G., Baigurmanova, D. and Jhunjhunwala, S. (2020). The new landscape of financial markets: Participants and asset classes. *Available at SSRN 4104690.*

appeared as the "kid brother" of sophisticated modelling paradigms. Then at the beginning of new millennium data entered the modeling scene and started gaining importance. In the next phase, it appeared as "data driven decision making". This was then followed by the era of "big data". Soon, thereafter, "big data" became the "big brother" of decision modeling. We at OptiRisk have kept abreast of the emergence of big data, data analytics and data science and have observed their growing importance in financial decision making. Today data is "the king" the preponderance of data in almost all aspects of Political, Economic, Social, and Technological in short (PEST) is tipping the scale in favour of "data" as opposed to models for decision making. This paradigm shift of [data and models]; from "kid brother" to "big brother" has ushered in the era of "Alternative Data". As early as 2017 Greenwich Associates' research report[5] "Alternative Data for Alpha" pointed out, "on the trading floors of the world's largest investment managers, the use of Alternative Data to make institutional investments is still in its infancy".

ORGANIZATION OF THE HANDBOOK

An Overview of Alternative Data is presented in **Chapter 1**. This chapter is a joint effort of the OptiRisk's editorial team together with Dr Keith Black and Dr Ganesh Mani of the FDPInstitute.

Dr Black and Dr Mani bring their knowledge of the finance industry to this chapter, namely, **Chapter 1- Alternative Data: Overview**. This chapter is followed by five parts.

In **Part I ALTERNATIVE DATA: PROCESSING AND IMPACT**, there are two chapters. **Chapter 2** is by David Jessop who presents his Contemplation and Reflection on using alternative data for trading and fund management. The second chapter, namely, **Chapter 3**: "Global Economy and Markets Sentiment Model" is by the research team of Northwestern Mutual. The authors [Jacob Gelfand, Kamilla Kasymova, Seamus M. O'Shea and Weijie Tan] introduce the novel data source "GDELT" and discuss its impact on financial market.

The **COUPLING** of **MODELS WITH ALTERNATIVE DATA FOR FINANCIAL ANALYTICS** is discussed in **Chapter 4** appearing in **Part II** of the Handbook. In the SenRisk project (see http://senrisk.eu/): a project funded by EU, a consortium of OptiRisk, Fraunhofer ITWM, and others investigated how "Macro News" can be exploited in sovereign and corporate bond modeling. AI, Machine Learning & Quantitative Models form the back-drop of research directed towards the application AI/ML and neo-classical Quant Models in Finance. In **Chapter 5**, we consider this topic and argue that these two modeling paradigms interact and are closely coupled.

[5]McPartland, K., 2017. *Alternative Data for Alpha*, s.l.: Greenwich Associates.

Our thesis is illustrated with an exemplar problem of predicting short-term market movement.

PART III HANDLING DIFFERENT ALTERNATIVE DATASETS

This part is made up of five chapters which cover handling different alternative datasets like news, micro-blog, company filings, earning calls and sensors data. The first two chapters discuss Asset Allocation Strategies Enhanced by Micro-blog and Enhanced by News. In the next two chapters the authors describe several methods to Extract Structured Datasets from Textual Sources like company filings and earning calls. The fifth chapter concern various kinds of Sensors Data.

How researchers have used News (meta) data, Sentiment (meta) data and other alternative data sources is presented in **Part III**. Two OptiRisk whitepapers may interest the readers as we first introduce the concept of filters, RSI filters to be specific. The (RSI) filters use only market data. We then proceed to describe two models in which these filters are enhanced using news (meta) data and micro-blog (meta) data respectively; see **Chapter 6** and **Chapter 7**, respectively.

PART IV ALTERNATIVE DATA USE CASES IN FINANCE

This part presents several case studies of applying alternative data in finance. We focus on three major applications of alternative data in finance, which include application in Trading and Fund Management to Find new Alpha, application in Risk Control and application in ESG.

Application in Trading and Fund Management (Finding new Alpha) includes three chapters:

- Media Sentiment Momentum: Global Equity Media-Driven Price Drift
- Defining Market States with Media Sentiment

Application in Risk Control includes three chapters:

- A Quantitative Metric for Corporate Sustainability
- Hot off the Press: Predicting Intraday Risk and Liquidity with News Analytics
- Exogenous Risks Alternative Data Implications for Strategic Asset Allocation—Multi-Subordination Levy Processes Approach

Case Studies on ESG there are three chapters:

- ESG Controversies and Stock Returns

- Oil and Gas Drilling Waste—A Material Externality
- ESG Scores and Price Momentum Are Compatible: Revisited

A list of alternative data providers is supplied in the last part of the Handbook:

PART V DIRECTORY OF ALTERNATIVE DATA VENDORS

In the directory fourteen (14) Alternative Data Vendors are listed. The information about these vendors and their products and services are available in the public domain and listed in their respective web portals. But having the information in a compact format and in one place provides a convenient and comparative summary.

Suggested Reading Sequence: After reading the overview chapter the rest of this book can be read in any order, depending on the reader's interest and job role, as the chapters have all been written independently. There is some referencing between the chapters but this does not hinder the understanding of content within each chapter. The purpose of this handbook is to invigorate and instigate readers to become active and pursue the exploration of sentiment analysis in finance. There are many applications to be explored, techniques and methods to be tested. Our mission is to motivate and excite the researchers and equally practitioners to participate in the research and development or exploitation of research knowledge in their respective areas of interest.

We have made a start in compiling the **Volume II** of the Handbook. We continue with the same theme and introduce (i) recent progress in theory and (ii) more use cases.

Gautam Mitra and Kieu Thi Hoang,
OptiRisk Systems Ltd, London

Acknowledgments

We first of all thank the members of the editorial panel for their painstaking effort in (i) the thorough review of the chapter contributions and (ii) their helpful discussions in determining the Parts and the Chapter structure of the Handbook.

We further thank our distinguished contributors who have given due care and effort in interacting with the reviewers and have revised their chapters in a timely fashion.

We thank Callum Fraser, the Editor of Mathematics Books CRC Press/Taylor & Francis Group, for his constructive suggestions in consummating the publication agreement. Further thanks are due to him and Mansi Kabra of the editorial team for working closely with us and speeding up the publication process.

Finally, we thank the members of our Handbook Team (Shradha Berry, Shrey Jhunjhunwala and Akshita Porwal), who have given their best and made the task of compilation of the Handbook go forward as clockwork.

Editors and Contributors

EDITORS

Gautam Mitra is founder and MD of Optirisk Systems. He is internationally renowned research scientist in the field of Operational Research in general and computational optimization and modeling in particular. He is an alumni of UCL and currently a visiting professor at UCL. In 2004 he was awarded the title of “distinguished professor” by Brunel University in recognition of his contributions in the domain of computational optimization, risk analytics and modeling. Professor Mitra is also the founder and chairman of the sister company UNICOM seminars.

Christina Erlwein-Sayer is a consultant and associate researcher at OptiRisk Systems. Her research interests lie in financial analytics, portfolio optimisation and risk management with sentiment analysis, involving time series modelling and machine learning techniques. She holds a professorship in Financial Mathematics and Statistics at HTW University of Applied Sciences, Berlin. She completed her PhD in Mathematics at Brunel University, London in 2008. She was then a researcher and consultant in the Financial Mathematics Department at Fraunhofer ITWM, Kaiserslautern, Germany. Between 2015 and 2018, prior to joining HTW Berlin in 2019, she was a full-time senior quantitative analyst and researcher at OptiRisk Systems, London, UK. She teaches modules on statistics, machine learning and financial mathematics and is part of the CSAF faculty. Christina is an experienced presenter at conferences and workshops: among others, she presented at workshops in London, IIM Calcutta in Kolkata and Mumbai and in Washington to World Bank.

Kieu Thi Hoang is a Financial Analyst and Relationship Manager at OptiRisk Systems. Kieu has a bachelor’s degree (with high distinction) in International Economics from Foreign Trade University, Hanoi, Vietnam. She was among the top 10% of all the global candidates in her CFA level 2 examination (December 2020). Kieu has a strong foundation in advanced financial analysis and work experience in the finance industry. She has years of experience working at different renowned BFSI firms in Vietnam. Joining OptiRisk Systems as a Financial Analyst and Relationship Manager, she has done a lot of thorough research on alternative data in company projects. She also works with a variety of alternative data providers who are partners of her firm.

Diana Roman is a Consultant and Research Associate at OptiRisk Systems. After completing her PhD at Brunel University under late Professor Darby-Dowman and Professor Mitra, Dr Roman joined OptiRisk Systems as a software developer. She had designed the scenario-generator library which was used inSPInEthe first version of the SP Tool developed by OptiRisk Systems. Together with Professor Mitra she has written a few seminal papers on the topic of portfolio construction with downside risk control in general and use of Second Order Stochastic Dominance (SSD) in particular. Dr Roman is a senior lecturer in the Department of Mathematics at Brunel University London.

Zryan Sadik is a senior Quantitative Analyst and Researcher at OptiRisk Systems. Dr Sadik has a bachelor's degree in Mathematics from Salahaddin University – Erbil in the Kurdistan region of Iraq. After working as an IT technician, he pursued an MSc Degree in Computational Mathematics with Modelling at Brunel University, London (2012). Dr Sadik completed his PhD in Applied Mathematics with a thesis on the "Asset Price and Volatility Forecasting Using News Sentiment" at Brunel University, London (2018). His research interests include news sentiment analysis, macroeconomic sentiment analysis, stochastic volatility models, filtering in linear and nonlinear time series applying Kalman filters, volatility forecasting as well as optimization and risk assessment. His current research interests lie in the areas of empirical finance and quantitative methods, and the role of Alternative data in financial markets. He has been involved in developing predictive models of sentiment analysis, and sentiment-based trading strategies for the last seven years. These models and strategies are developed in C, C++, MATLAB, Python and R as appropriate. His prior studies include the impact of macroeconomic news on the spot and futures prices of crude oil, and the impact of firm-specific news on the movement of asset prices and on the volatility of asset price returns. Dr Sadik is fluent in Kurdish (his native language), as well as in English and Arabic.

EDITORIAL PANEL

Keith Black is the managing director and program director of the FDP Institute. Previously, he served as the managing director of content strategy at the CAIA Association, where he was a co-author of the CAIA curriculum. During a prior role, Keith advised foundations, endowments and pension funds on their asset allocation and manager selection strategies in hedge funds, commodities and managed futures. He has also traded commodity and equity derivatives and built quantitative stock selection models. Dr. Black earned a BA from Whittier College, an MBA from Carnegie Mellon University and a PhD from the Illinois Institute of Technology. He is a CFA, CAIA and FDP charterholder.

Arkaja Chakraverty is a Senior Research Associate at OptiRisk Systems. Arkaja is a dynamic academic actively engaged in research in the domain of corporate finance and financial markets. She has the experience of investing in Indian equity derivatives, namely, futures and options. Through a number of consulting projects involving small-scale enterprises (start ups companies such as — Man Capital, Clark & Kent Inc), she had been active in the industry. She received her PhD from Indian School of Business in Financial Economics in 2017; she is affiliated with Higher School of Economics, Moscow. Currently she is based out of Melbourne and is working on a series of research papers. As of April 2021, Arkaja has joined the OptiRisk team as an Associate (Consultant).

Yuanqi Chu joined OptiRisk Systems as an industry sponsored PhD student in the Department of Mathematics at Brunel University. Yuanqi is supervised by Professor Keming Yu and her research interests lie primarily in developing new methods on regression analysis of skewness, over-dispersion and heavy-tailedness. She obtained her master's degree in Statistics (with Distinction) from University of Nottingham in 2019. Yuanqi has a strong programming background and is well versed in R, MATLAB and SPSS.

Alexander Gladilin is the CEO of a start-up company Transolved and a research associate of OptiRisk. He is also a Doctoral Researcher at the Mathematics Department of Brunel University, actively involved in the research projects funded by OptiRisk. Alexander has work experience and master's degrees in finance and Data Analytics. His research focuses on applications of implied volatility and alternative data in financial modelling as applied to the BFSI sector.

Ganesh Mani is on the adjunct faculty of Carnegie Mellon University and is considered a global thought leader, helping scale human expertise via AI in many application areas, particularly financial services. He has been a pioneer in applying innovative techniques to multiple asset classes and has worked with large investment management firms incl. hedge funds, after having sold one of the earliest AI/ML-based asset 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". He has an MBA in Finance and a PhD specializing 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 often invited to speak and moderate panels at industry conferences around “The Art of Data Science” theme.

CONTRIBUTORS

Shradha Berry joined OptiRisk Systems in October 2019 as a Data Scientist and Research Analyst. She has over 5 years' experience in software development with a multinational consulting company in India. Shradha has a Bachelor's degree in Information Technology from Meghnad Saha Institute of technology, Kolkata, India and a Master's degree (with Distinction) in Data Science and Analytics from Brunel University, London. In OptiRisk her research was focused on AI & ML applied to quant finance models for the BFSI sector. Shradha left OptiRisk in June 2022.

Zhixin Cai was a Quantitative Research Intern at OptiRisk from 2018 to 2019. He has a Bachelor's Degree in Economics and Finance in The College of Economics, Henan University, China. He engaged in the exchange programme to Cardiff University from 2014 to 2016 and has done the BSc Econ Economics and Finance and BSc Econ Banking and Finance with first class result. He also holds a MSc Financial Risk Management in University College London.

Matteo Campellone is the founder and Executive Chairman of Brain, a company focused on the development of algorithms for trading strategies and investment decisions. He holds a Ph.D. in Physics and a Master in Business Administration. Matteo's past activities included Financial Modelling for financial institutions and Corporate Risk and Value Based Management for industrial companies. As a Theoretical Physicist he worked in the field of statistical mechanics of complex systems and of non-linear stochastic equations.

Joshua Clark-Bell is a Quantitative Researcher and has been working at MarketPsych for over a year. He holds a degree in econometrics and has an interest in applications of NLP towards sustainable finance.

Francesco Cricchio is the founder and Chief Executive Officer of Brain, a company focused on the development of algorithms for trading strategies and investment decisions. Francesco obtained his Ph.D. in Computational Methods applied to Quantum Physics from Uppsala University in 2010. He focused his career in solving complex computational problems in different sectors using a wide range of techniques, from density functional theory in Solid State

Physics to the application of Machine Learning in different industrial sectors with focus on Finance.

Giuliano De Rossi is the head of Portfolio Analytics in Prime Services at Goldman Sachs. Prior to this, he worked at Macquarie and PIMCO. He also spent six years in the Quant research team at UBS. Giuliano has a PhD in economics from Cambridge University and worked for three years as a college lecturer in economics at Cambridge before joining the finance industry on a full-time basis. Giuliano's Masters degree is from the LSE and his first degree is from Bocconi University in Milan. He has worked on a wide range of topics, including pairs trading, low volatility, the tracking error of global ETFs, cross asset strategies, downside risk and applications of machine learning to finance. His academic research has been published in the Journal of Econometrics and the Journal of Empirical Finance.

Dan diBartolomeo is President and founder of Northfield Information Services, Inc. Based in Boston since 1986, Northfield develops quantitative models of financial markets. He sits on boards of numerous industry organizations include IAQF and CQA, and is a director and past president of the Boston Economic Club. His publication record includes fifty books, book chapters and research journal articles. In addition, Dan spent several years as a Visiting Professor at Brunel University and has been admitted as an expert witness in litigation matters regarding investment management practices and derivatives in both US federal and state courts. He became editor in chief of the *Journal of Asset Management* at the start of 2019.

Jacob Gelfand, CFA, Director of Quantitative Strategy and Research, Investment Risk Management, is responsible for macro research and analysis for the fixed income, currencies, equity and cross-asset class portfolios. His focus areas include global markets risk analysis, asset allocation, investment style and risk management, relative value analysis and other quantitative and global macroaspects of risk and portfolio management.

Mr. Gelfand joined Northwestern Mutual in 2004 as a member of the Fixed Income Department at Mason Street Advisors (MSA). In 2015, he joined the Investment Strategy Division and in 2016 became a member of the Investment Risk Management Division. Before joining Northwestern Mutual, Mr. Gelfand was a Senior Consultant with Cap Gemini, serving clients in investment management and financial industries. Mr. Gelfand is an adjunct faculty in the Lubar School of Business, University of Wisconsin- Milwaukee, where he teaches classes on Global Investments, and Options and Derivatives. Mr. Gelfand serves on the board of the CFA Society in Milwaukee, Wisconsin.

Mr. Gelfand received a MS in Computer Science with Honors from the Moscow State University of Civil Engineering (MSUCE), Russia, and an MBA in Finance and Strategy with Honors from the University of Chicago Booth School of Business. He holds the Chartered Financial Analyst (CFA) designation.

Ryoko Ito works on equities execution research and content generation at Goldman Sachs. Prior to this, she worked at UBS, during which she managed a team of quants responsible for designing and running econometric models used in firm-wide regulatory stress tests. She was also a postdoctoral researcher at Oxford University, during which she lectured financial econometrics to graduate students at Saïd Business School. She has published academic papers on time series analysis and financial econometrics. Her recent work on modeling time series with zero-valued observations appeared in the Journal of Econometrics in 2019. Ryoko has a PhD in economics from Cambridge University. She undertook research projects with her PhD sponsors; one on liquidity prediction and order-book analysis in high-frequency FX at Morgan Stanley, and another on modeling systemic risk in the European banking system at the International Monetary Fund. Ryoko also completed Part III of the Mathematical Tripos at Cambridge University.

David Jessop is Head of Investment Risk at Columbia Threadneedle Investments EMEA APAC. He has responsibility for overseeing the independent investment risk management process for all portfolios managed in the EMEA region.

Before joining Columbia Threadneedle Investments, David was the Global Head of Quantitative Research at UBS. Over his 17 years at UBS, his research covered many topics but in particular he concentrated on risk analysis, portfolio construction and more recently cross asset factor investing / the application of machine learning and Bayesian techniques in investment management. Prior to this he was Head of Quantitative Marketing at Citigroup. David started his career at Morgan Grenfell, initially as a derivative analyst and then as a quantitative portfolio manager.

David has an MA in Mathematics from Trinity College, Cambridge.

Dan Joldzic, CFA, FRM is CEO of Alexandria Technology, Inc, which develops artificial intelligence to analyze financial news. Prior to joining Alexandria, Dan served dual roles as an equity portfolio manager and quantitative research analyst at Alliance Bernstein where he performed factor research to enhance the performance of equity portfolios.

Christopher Kantos is a Managing Director at Alexandria Technology. In this role, he focuses on maintaining and growing new business in EMEA and also co-heads research efforts at Alexandria, focusing on exploring ways in which natural language processing and machine learning can be applied in the financial domain. Prior, he spent 15 years working in financial risk at Northfield Information Services as a Director and Senior Equity Risk Analyst. Mr. Kantos earned a BS in computer engineering from Tufts University.

Kamilla Kasymova is a Quantitative Research and Analytics Director in the Investment Risk Management of Northwestern Mutual Life Insurance Company. She joined the company in 2014. Kamilla has extensive experience using analytical and research techniques (time series analysis and forecasting, econometrics, financial mathematics) in developing internal capital market assumptions and economic scenario generator used by Actuarial and Investment departments across the company. Prior to joining Northwestern Mutual, Kamilla taught various undergraduate mathematics and economics classes during her graduate studies.

Kamilla has a BSc in Economics from Moscow State University, MSc in Finance from University of Ulm, MSc in Mathematics from University of Wisconsin-Milwaukee and PhD in Economics from University of Wisconsin-Milwaukee.

Richard L. Peterson is CEO of MarketPsych Data which produces psychological and macroeconomic data derived from text analytics of news and social media. MarketPsych's data is consumed by the world's largest hedge funds. Dr. Peterson is an award-winning financial writer, an associate editor of the Journal of Behavioral Finance, has published widely in academia and performed postdoctoral neuroeconomics research at Stanford University.

Changjie Liu is Head of Research at MarketPsych. He has more than fifteen years of experience in investigating sentiment data, including ten years in MarketPsych. He leads the quant and web development team in planning and developing sentiment-based investment strategies and products. He holds a MFE from UC Berkeley and a Computer Science BSc with honors from NUS.

Anthony Luciani, MSc is a Quantitative Researcher at MarketPsych. He constructs and validates sentiment-based signals, with a focus on relative value strategies and using machine learning techniques. His research is built upon an education in mathematics and information theory, and his prior work in behavioral financial studies.

Seamus M. O'Shea, CFA is the Managing Director of the Public Investment in Northwestern Mutual. He joined Northwestern Mutual in 2018 after spending two years with a specialty finance start-up based outside of Boston. In his current role, he is responsible for the sector allocation process for the \$105bn+ Public Markets fixed income portfolio, which includes corporate, mortgage-backed and municipal bonds. In addition, he has direct portfolio management oversight for the internally managed EM assets.

Prior to trying his hand in the start-up world, he co-managed \$50bn in U.S. Credit-oriented separate accounts and mutual funds at Wellington Management, where he was a Managing Director and the lead risk-taker for the Emerging Markets, High Yield and ABS segments of client accounts. In addition to his portfolio management duties there, he maintained extensive client-facing responsibilities and was regularly sought out to present to current and prospective clients and consultancies both in the US as well as overseas.

Prior to joining Wellington, Seamus was a Global Bond portfolio manager at Pacific Investment Management Company (PIMCO) in Newport Beach, CA, where he helped manage over \$60bn in Global-Agg benchmarked accounts. He has 15+ years of investment experience, and began his career at Fidelity Investments.

Seamus holds a MBA from MIT's Sloan School of Management and a BA in Mathematics from the College of the Holy Cross. He is a CFA Charterholder, and a citizen of the United States, Canada and Ireland.

Matus Padysak has a strong background in probability and statistics of which he holds a master´s degree. Currently, he is a Ph.D. candidate with a research interest in pension funds investments. Apart from the research, his responsibilities also lie in teaching several undergraduate courses. He is always keen to utilize all his knowledge in finances, investing or trading. To escape from the research and clear the head, Matus likes to swim, bike and run.

Tiago Quevedo Teodoro is a Quantitative Researcher at MarketPsych. He specialises in applying statistical and machine learning methods toward sentiment usage in timing and event studies. He holds a Ph.D. in Physical Chemistry as well as an MSc in Finance & Technology, with 16 peer-reviewed publications in top scientific journals.

Boryana Racheva-Iotova is Senior Vice President, Senior Director of Buy-Side Strategy at Factset, driving the strategy and vision for FactSets's solutions for institutional asset managers and asset owners. Prior to that, she led the Risk and Quantitative Analytics Business Line for FactSet, responsible for all risk and quantitative solutions, which includes quantitative research, product development and business strategy. She is a co-founder of

FinAnalytica and the former Global Head of Risk at BISAM and has over 20 years of experience in building risk and quant portfolio management software solutions. Before founding FinAnalytica, Ms. Racheva-Iotova led the implementation of a Monte-Carlo based VaR calculation to meet the Basel II requirements at SGZ Bank, as well as the development of six patented methodologies for FinAnalytica. In 2018, she received the Risk Professional of the Year Award from Waters Technology based on her achievements in building risk management software solutions and translating the latest academic advancements in practical applications to meet the needs of financial industry practitioners. Ms. Racheva-Iotova earned a Master of Science in Probability and Statistics at Sofia University and a Doctor of Science from Ludwig Maximilian University of Munich.

J. Blake Scott is the President of Waste Analytics LLC. Blake has been a leader in the oil and gas sector for more than 25 years. As founder and president of Scott Environmental Services and Scott Energy Technologies, Blake pioneered groundbreaking research and development in the drilling-waste recycling industry. Blake has received multiple awards and recognitions for his work, and has presented to organizations such as the Interstate Oil and Gas Compact Commission, has served on ITRC's Solidification/Stabilization Team and was an expert for GRI's Oil and Gas Sector Standard. Blake founded Waste Analytics in 2018 in order to provide unique data on oil and gas companies for current insight and financial prediction. Blake has a BA and MBA from Texas Christian University.

Michael Steliaros is global head of Quantitative Execution Services at Goldman Sachs. He is responsible for the research, development and implementation of quantitative processes for portfolio and electronic trading, as well as the management of the firm's relationships with the quantitative clientbase across regions. Michael manages a variety of teams globally, spanning algorithmic research, portfolio quants, client solutions, analytics and quantitative content generation. He joined the firm as a managing director in 2017.

Weijie Tan is Associate Quantitative Analyst in the Investment Risk Management of Northwestern Mutual. He joined Northwestern Mutual in 2020. In his current role, he focuses on Capital Market Assumptions, which produces long-term return, volatility and correlation projection of a variety of asset classes. Besides, he builds and maintains indicators based on sentiment analysis to support asset allocation. Weijie holds a Ph.D. in Engineering from the University of Wisconsin - Milwaukee and a bachelor's degree in Economics from the Shanghai University of Finance and Economics.

Ziwen Tan was a Quantitative Research Intern at OptiRisk from 2018 to 2019. Before that, she was a Research Assistant for two years at the University of Toronto. She became a Senior Data Analyst at PHD Canada in 2020. Ziwen holds a Bachelor's Degree with Distinction in Statistics and Linguistics from the University of Toronto. She is currently pursuing her Master's Degree in Library and Information Science at the University of Illinois Urbana-Champaign.

Gareth Williams is a Data Scientist at Transolved Ltd. He had previously completed an undergraduate degree in Physics with Mathematics at Keele University before obtaining a PhD in Theoretical Physics at the University of Liverpool. His research interests include applied Bayesian statistics and computational physics, in particular the calculation of marginalized likelihoods and the use of the Bayes factor as a tool for model comparison.

William Zieff is Director at Northfield Information Services, a leading provider of financial risk analytics. He has extensive experience in finance and investing. Prior to Northfield, he was Chief Investment Officer of Global Strategic Products, Wells Fargo Asset Management; Partner and Co-Chief Investment Officer of Global Asset Allocation at Putnam Investments; and Director of Asset Allocation at Grantham, Mayo, Van Otterloo. He has taught master's and undergraduate courses in finance at Northeastern University and Boston College. He holds an A.B. degree in Economics and Mathematics from Brown University and MBA from Harvard Business School.

Abbreviations

ABBREVIATION	EXPLANATION
ACOG	Association of Central Oklahoma Governments
ADF	Augmented Dickey–Fuller test
AHT	After-Hours Trading
AGG	Aggregated bond index ETF
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
API	American Petroleum Institute
API	Application Programming Interface
AQ	Analysts' Questions
ARCH	Auto-regressive Conditional Heteroskedastic
ARIMAX	Autoregressive–Moving-Average Model with Exogenous Inputs Model
AUM	Assets under Management
BERT	Bidirectional Encoder Representations from Transformers
BLMCF	Brain Language Metrics on Company Filings
BOW	Bag of Words
BSI	Brain Sentiment Indicator
CAC	Central America and the Caribbean
CAGR	Compound Annual Growth Rate
CAPM	capital asset pricing model
CDS	Credit Default Swap
CFP	Corporate Financial performance
CIS	Commonwealth of Independent States
CLOB	Central Limit Order Book
CNN	Convolutional Neural Network
DM	Diebold-Mariano
DNS	Domain Name System
DRSI	Derived RSI
ECG	Electric Company of Ghana
ECT	Earnings Calls Transcripts
EEL	Estimated Environmental Liability

EMA	Exponential Moving Average
EMD	empirical modal decomposition
EMH	Efficient Market Hypothesis
ESG	Environmental, Social and Governance
ESPP	Employee Stock Purchase Plan
ETL	Expected Tail Loss
EU	European Union
EVSI	External Vulnerability Sentiment Index
FAA	Federal Aviation Administration
FASB	Financial Accounting Standards Board
FN	False-Negative
FP	False-Positive
FSD	First-Order Stochastic Dominance
FT	Financial Times
FTRI	Financial Technology Research Institute
FX	Foreign Exchange
GARCH	Generalized ARCH
GCAM	Global Content Analysis Measures
GDELT	Global Database of Events, Location and Tone
GDPR	Global Data Protection Regulation
GEMS	Global Economy and Markets Sentiment
GHG	Greenhouse Gas Emissions
GMM	Gaussian mixture model
GPS	Global Positioning System
GRI	Global Reporting Initiative
GUI	Graphical User Interface
HFT	High-Frequency Trading
ICBDAC	International Conference on Big Data Analytics and Computational
IG	Investment Grade
IIRC	International Integrated Reporting Council
IMF	International Monetary Fund
IoT	Internet of Things
IPIECA	International Petroleum Industry Environmental Conservation Association
IRR	Internal Rate of Return
IRS	Interest Rate Swap
IS	Impact Scores
IVSI	Internal Vulnerability Sentiment Index
LM	Loughran McDonald
LP	Linear Programming
LR	logistic regression

LSTM	long short-term memory
LT VSI	Long-Term Vulnerability Sentiment Index
MA	Moving Average
MAAQ	Management Answers to Analysts' Questions
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error
MASE	Mean Absolute Scaled Error
MD	Management Discussion
MIP	Mixed Integer Programming
ML	Machine Learning
MLE	maximum likelihood estimation
MOM	Momentum
MPT	Modern Portfolio Theory
MRG	Mergers and Acquisitions
MRSI	Micro-blog Relative Strength Index
NCQ	Neo-Classsical Quant
NETL	National Energy Technology Lab
NIH	National Institutes of Health
NLP	Natural Language Processing
NORM	Naturally Occurring Radioactive Material
NRSI	News Relative Strength Index
NYSE	New York Stock Exchange
OCC	Oklahoma Corporation Commission
OHLC	Open, High, Low and Close
OLS	ordinary least square
PADEP	Pennsylvania Department of Protection
PAH	Polycyclic Aromatic Hydrocarbons
PEAD	Post-Earnings Announcement Drift
PII	Personally Identifiable Information
QA	Questions and Answers
RCRA	Resource Conservation and Recovery Act
RF	Random Forest
RFF	Resources for the Future
RFQ	Request for Quote
RFR	Risk Free Rate
RIC	Reuters Instrument Code
RMA	Refinitiv MarketPsych Analytics
RM-ESG	Refinitiv MarketPsych ESG
RMSE	Root Mean Square Error
RND	Risk Neutral Density
ROP	Rate of Penetration
RPNA	RavenPack News Analytics

RSI	Relative Strength Index
SASB	Sustainability Accounting Standards Board
SCOTUS	Supreme Court of the United States
SD	Stochastic Dominance
SDG	Sustainable Development Goal
SEC	Securities and Exchange Commission
SPE	Society of Petroleum Engineers
SSD	Second Order Stochastic Dominance
SSE	Shanghai Stock Exchange
ST VSI	Short-Term Vulnerability Sentiment Index
SVM	support vector machine
TB	Terabytes
TN	True-Negative
TP	True-Positive
TRMA	Thomson Reuters MarketPsych Analytics
TSD	Third-order Stochastic Dominance
UAT	User Acceptance Training
UK	United Kingdom
UN	United Nations
UNPRI	United Nations Principles on Responsible Investing
USEPA	United States Environmental Protection Agency
USSEC	United States Securities and Exchange Commission
UTC	Coordinated Universal Time
VIX	Volatility Index
VSI	Vulnerability Sentiment Indices
VSW	Velocity Solar Wind
XGB	XGBoost eXtreme Gradient Boosting

CHAPTER 1

Alternative Data: Overview

Gautam Mitra

Research Director, OptiRisk Systems and UCL Department of Computer Science, London, United Kingdom

Kieu Thi Hoang

Senior Financial Analyst, OptiRisk Systems

Alexander Gladilin

Research Associate, OptiRisk Systems

Yuanqi Chu

Sponsored PhD Candidate and Intern, OptiRisk Systems

Keith Black

Managing Director and Program Director, FDP Institute

Ganesh Mani

Adjunct Faculty, Carnegie Mellon University

CONTENTS

DOI: 10.1201/9781003293644-1

1.1 INTRODUCTION

The finance industry has always employed data. The digital era has generated exponential growth in various types of data, and the big data phenomenon has revolutionized the modern age. A growing number of researchers and investors have embraced big data advances to enhance investment performance. What set the current evolution apart are large volumes of alternative data, which have the potential to profoundly reshape financial landscape in the foreseeable future. Alternative data can be conceptualized as the data derived from outside of the standard repository of financial data. [Quinlan and Cheng, 2017] refer to the alternative data sources as "any useable information or data that is not from a financial statement or report". Different from traditional repertoire of financial information, such datasets are granular, real time and speed centric, including transaction activities, social-media streams, satellite images, clickstream data, digital footprints and other emerging data sensed via the Internet of Things (IoT).

The exploitation of alternative data has been catalyzed by the boom of the data vending industry [Kolanovic and Krishnamachari, 2017] and the belief that integrating alternative data can reduce the cost of obtaining information [Goldfarb and Tucker, 2019]. The application of alternative data has attracted great interest in various perspectives within today's financial landscape, including trading strategy generation [Xiao and Chen, 2018], enhancing credit scoring [Cheney, 2008]; [Djeundje et al., 2021], portfolio construction and allocation [Henriksson et al., 2019]; [Bertolotti, 2020]; [De Spiegeleer et al., 2021], among others. Using contemporary data science and Artificial Intelligence (AI) techniques, economic researchers and market practitioners can exploit key alternative data to interpret and condense increasing amounts of traditional information, extracting pithy insights for cost reduction, revenue enhancement and maintaining a competitive edge.

While the intertwinement between alternative data and finance systems is broadening the possibilities for innovation and value creation, various challenges and issues also arise through the exploration of the data. As [Dourish and Gomez Cruz, 2018] point out, "Data must be narrated – put to work in particular contexts, sunk into narratives that give them shape and meaning, and mobilized as part of broader processes of interpretation and meaning-making", the lack of an eligible narration and characterization paradigm makes alternative data employment more vulnerable to issues like data ownership, security and privacy than traditional data sources. The challenges and potential pitfalls of managing, using and analyzing alternative data sources should be highlighted by researchers and financial practitioners in the process of extracting improved analytics-driven insights.

In this chapter we describe the current landscape of alternative data with a view to provide sufficient information and guidance for analysts to exploit such datasets. The rest of this chapter is organized as follows. In Section 1.2 we discuss the increasing availability of data in the financial arena. In Section 1.3 we outline the role and use of traditional data. In Section 1.4 we formally define alternative data and highlight the differences between traditional and alternative data. In this section we bring out the differences between alternative data and big data and briefly discuss a few use cases and consider some ethical challenges. In Section 1.5 we consider the phenomenal growth of alternative data; our discussions cover recent history, the drivers of growth and the positive impact experienced in the financial markets. Alternative data providers collect, aggregate, structure and act as a financial intermediary for the buyers of alternative data. In Section 1.6 we describe the services provided by these vendors; some examples of established data vendors are cited and their services are discussed. In Section 1.7 we consider the challenges faced by diverse users and financial market participants who want to exploit alternative data and we end the chapter with a summary discussion and conclusion in Section 1.8.

1.2 DATA

In recent times in the realm of analytics, data has gained substantial importance. The growth of data analytics and data science has vindicated this; [Donoho et al., 2000] have articulated well via: "The coming century is surely the century of data". Recent technological innovations are based on this central role of data and the proliferation of globally produced data. There has been an explosion in the volume, velocity and variety of the data which are collected by scraping web portals; and from mobile phone usage, social media activities, online payment records, customer service records and embedded sensors. As the OECD [The Organization for Economic Cooperation and Development, 2013] suggests, "In business, the exploitation of data, promises to create added value in a variety of operations, ranging from optimizing the value chain, manufacturing production to an efficient use of labor and improved customer relationships". The technologies are constantly evolving; they intertwine and embrace "Big Data", the "Internet of Things" (IoT) and the "Internet of Signs" [O'Leary, 2013]. These achieve the desirable goals of knowledge management, knowledge sharing and effective decision-making.

According to [Aldridge, 2015] a decade ago, finance was a small-data discipline. Because data was a scarce resource it naturally relied upon small-data paradigm. To most investors, exchanges offered only four prices per stock per day: Open, High, Low and Close (OHLC). Data at higher frequency was not stored by even the largest market makers [Aldridge, 2015]. [Wang, 2008] reinforces this paradigm shift and observes: "Financial econometrics is only made possible by the availability of vast economic and financial data". In the new landscape of the financial markets, see [Mitra et al., 2020], data takes the central role and has led to the transformation of financial service sectors and institutions. In discussing applications of "Big Data in Finance", [Goldstein et al., 2021] recently pointed out that three properties, namely, large size, high dimension and complex structure which taken together underpin the role of "Big Data in Finance" research.

"According to IBM, companies have captured more data in the last two years than in the previous 2000 years" [Syed et al., 2013]. "ONE MEASURE OF PROGRESS in empirical econometrics is the frequency of data used" [Engle, 2000]. In contrast to the availability of OHLC data reported the following day (on the T+1 basis) in the small-data age, in current financial markets, a stock can experience 500 quote changes and 150 trades in one microsecond [Lewis, 2014]. High-frequency trading (HFT) has grown tremendously and is becoming increasingly important in the functioning of the financial markets. In the world of "Algo Trading", massive amounts of high-frequency data are consumed by sophisticated algorithms. Algo Traders take advantage of HFT to spot profitable opportunities that may only be open for milliseconds.

In financial portfolio construction and optimization, high dimensionality is encountered when asset managers estimate the covariance matrix or its inverse matrix of the returns of a large pool of assets [Fan et al., 2011]. High-dimensional time series such as the country-level quarterly macroeconomic data [Stock and Watson, 2009], simultaneously observed return series for a large number of stocks [Barigozzi and Hallin, 2017] and large series of company financials based on their quarterly reports [Wang et al., 2019] also abound. These data sources have gained considerable attention in modern finance and economics. The "curse of dimensionality" that arises in financial data, presents new challenges from a computational and statistical perspective. As for financial market participants, they have to delve into real-time analytics and decision making to effectively exploit the value hidden in datasets.

Another important characteristic of finance-related big data is complexity. Unstructured data refers to information that is not organized and does not fall into a pre-determined model [Fang and Zhang, 2016]. So the data that is gathered from various sources, such as news articles, blogs, forums, twitters, emails or images, audio, video, is unstructured. Unstructured data is often characterized by high dimensionality [Goldstein et al., 2021]. Recent research findings have shed light on the prominent role of social media data in the analysis of financial dynamics.

From the era of Big Data, we have now moved on to alternative data ("alt-data"). Alternative data sources can be characterized as "any useable information or data that is not from a financial statement or report" [Quinlan and Cheng, 2017]. These are different from traditional sources of financial information; typically, asset prices, corporate annual reports and consumer spending. Such datasets are granular, real time and speed centric, including transactional activities, social-media streams, satellite images, clickstream data, digital footprints and other emerging data from the IoT. In financial landscape, market participants are embracing and unlocking alternative data to extract competitive edges and identify trading opportunities: [Jame et al., 2016] assert the value of crowdsourced forecasts in forecasting earnings and measuring the market's expectations of earnings; [Cong et al., 2021] highlight the promising utilization of different categories of alt-data in economics and business-related fields. [In et al., 2019] show the environmental, social and governance (ESG) data, when it is of high quality, is capable of mapping onto the investment decision-making processes. [Monk et al., 2019] emphasize that with efficient organizing and processing, alt-data can be employed to delve deep and gain insights of risk and generate operational alpha, [Berg et al., 2020] suggest that even the simple, easily accessible users' digital footprints are highly informative in predicting consumer default, to name only a few.

Today alternative data has assumed a central role in financial analytics and thereby in financial markets; it is well set to drive future innovations in financial systems. We observe that many challenges such as data acquisition, information extraction, analytical technique as well as data security and privacy have emerged and need to be suitably addressed through research and innovation.

1.3 TRADITIONAL DATA

Traditional data has been an input into quantitative investment decision making for over 40 years. In most countries, publicly traded companies make substantial disclosures on a quarterly basis. In addition, stock exchanges publish data on stock prices and trading volume, while investment banks and brokerage firms publish estimates for futures earnings releases.

Analysis of traditional investment information has focused on ratio-based measures from data disclosed on quarterly corporate income statements and balance sheets. Using the income statement, growth-oriented investors focus on revenue and earnings growth measures, comparing the current quarter and current year revenue and earnings relative to prior periods. Value investors use the market capitalization and the stock price of the company to analyze the cheapness of a company relative to income statement measures such as price/cash flow, price/earnings, or price/revenues or balance sheet measures such as price/book value of equity. Finally, stock prices can move substantially based on the concept of earnings or revenue surprise, where the stock price is expected to rise following a positive surprise where reported earnings or revenues exceeds the earnings or revenue expectation of the sell-side analysts.

Traditional data has been readily available and relatively straight forward to work with. Quantitative income statement and balance sheet data can be downloaded as panel datasets, with the reports of thousands of companies held in a structured database holding quarterly reports over historical time periods. Stock prices and analyst revenue and earnings estimates can be easily added, as they have been previously organized or tagged with the stock ticker by the data vendor.

Once the raw data is available, quantitative analysts can run backtests of this data, testing the historical response of stock prices to some combination of value and growth factors, as well as earnings or revenue surprises. The goal, of course, is to build a model where the combination of factors is predictive of future stock prices.

The academic literature on return anomalies documents a number of traditional quantitative factors that have predicted excess returns over long periods of time. Note that while these factors may earn excess returns relative to the market over long periods of time, they may also lag the market for significant periods of time. For example, while value stocks have outperformed growth stocks from 1980 to 1994, value underperformed from 1995 to 1999. More recently, value stocks beat growth stocks from 2000 to 2016, while growth triumphed from 2017 to 2020. Perhaps, value stocks outperform in weak stock markets when risk aversion is high and growth stocks outperform in strong stock markets when investors are risk seeking.

Fama and French pioneered studies on the returns to small cap and value stocks [Fama and French, 1992]. Value stocks track companies where the stock price is low relative to some measure of earnings or value, such as price-to-earnings, price-to-book, price-to-sales or price-to-cash-flow. For each dollar of stock price, value investors are buying more earnings, book value, revenues or cash flow than can be obtained by buying growth stocks. However, value stocks tend to have lower expected growth in these metrics compared to growth stocks.

Similarly, stocks with smaller market capitalizations have historically outperformed stocks with larger capitalizations over long periods of time. In the US, small stocks outperformed large stocks from 1981 to 1997 with a brief benefit for growth in 1990, while large stocks benefited from outperformance in 1998-1999. Small stocks once again outperformed from 2000 to 2018, while larger stocks had higher returns in 2019–2020.

Excess returns to stocks with price momentum were documented by Carhart [Carhart, 1997] as well as Jagadeesh and Titman [Jegadeesh and Titman, 1993]. Stocks that have outperformed the market over the trailing six to twelve months are likely to outperform over the subsequent six to twelve months. Conversely, stocks that have underperformed over a similar period are likely to continue to underperform. Note that the effect is the opposite in the short run, as stocks that have outperformed over the last month tend to underperform over the next few weeks.

Lawrence Brown pioneered the research [Brown, 1997] on earnings momentum and earnings surprise. Due to positive momentum in a company's business, positive news regarding earnings often lasts longer than one quarter. If markets were highly efficient, a positive earnings surprise would lead to an immediate and complete rise in the stock price to reflect this new information. However, a post-earnings-announcement drift has been documented, where investors could buy stocks after the earnings surprise is announced and still be able to outperform the market. Earnings surprises may be predicted by noting that analysts are increasing their earnings estimates during a calendar quarter.

Research from Richard Sloan has focused on quality stocks and accounting accruals [Sloan, 1996]. Investors focusing on quality of earnings seek to buy stocks where net income and earnings are highly correlated. That is, when companies report high quality profits that are received mostly in cash, there is a lower probability that earnings will subsequently be restated. When accounting accruals are high, net income is elevated relative to the cash flow received by companies with higher earnings quality. Companies that report higher net income by deferring expenses and accelerating revenues tend to find it more difficult to sustain the level of reported earnings growth.

There is also a strong signal in the change in shares outstanding of a company. Stocks that are reducing shares outstanding through buyback programs tend to outperform, as their purchase of stock increases the demand for the stock and reduces the supply. Earnings per share is defined as net income divided by shares outstanding. As the number of shares outstanding is reduced, earnings per share rises, even when net income is unchanged. Stocks that are increasing the number of shares outstanding tend to underperform, whether that issuance comes through employee stock options, secondary sales or stock swap mergers. Investors may also track the purchases and sales of stock by insiders, such as the CEO, CFO and board members. When insiders are buying shares, the company may outperform. Conversely, when a large number of insiders are selling a large portion of their shares, the company may underperform. Note that insider trading can be a signal for the entire stock market, where a sharp increase in insider sales across the entire market may predict a decline in the stock market index.

Modeling these stock market anomalies is relatively straight forward, as the data was available in a structured manner, mapped by dates, stock tickers and columns of fundamental data for each stock. Backtesting engines and various datasets are readily available from commercial vendors. There were two downsides, though, to this use of traditional data. First, because the data was readily available and highly structured, large numbers of investors had access to the same data and often built trading strategies using highly correlated signals. That is, many investors derived the same models based on the limited number of factors and the rich academic literature just explained. Second, the data was not very granular, as most of the new information came only when each company in the database reported their earnings, income statement and balance sheet on a quarterly basis. While stock prices and analyst estimates could change on a daily basis, most of the data disclosed by companies was available only four times each year. Because of the small number of data points on each company, models using traditional data are most useful for time periods of one to six months or longer. That is, the main value added is at the time of the next quarterly report, as many stocks have the majority of their annual idiosyncratic price movements during the one week of each calendar quarter when earnings are reported.

While users of this traditional data generally waited until earnings announcements to earn their excess returns, investors deploying alternative data may be able to profit over a shorter time period. While the holding period of traditional models may be denominated in months, models deploying alternative data may have holding periods of days to weeks. That is, while traditional models may profit from holding a stock after an earnings announcement, alternative data models may profit by predicting that earnings announcement.

1.4 ALTERNATIVE DATA

In contrast to the highly structured world of most traditional data sources, alternative data is unstructured, more messy, high volume and high frequency. Alternative data does not come from a single source already organized by stock ticker, as was the case of traditional data such as income statements, balance sheets, earnings estimates or stock prices.

The goal of alternative data is to build a forecast of corporate revenues and earnings more quickly than can be accomplished using the quarterly data disclosed by publicly traded companies. Alternative data comes in many forms, some of which are available in real time. While traditional data is available for all companies and is available in easily accessible database formats, many of the alternative data sources focus on a more narrow sector of companies. That is, every public company reports income statement and balance sheet data, while only healthcare companies report drug reactions to regulators and motor vehicle bureaus only report on the sales of automobile manufacturers.

A very promising avenue of exploration for alternative data is to derive the revenue of privately held companies. In the US, there are 3,000 public companies with revenues over 100 million dollars, while there are over 30,000 private companies with similar revenues. Due to regulatory requirements, public companies offer regular transparency regarding revenues and profits. Private companies, however, are considered more difficult to analyze, as they do not have the required disclosures of public companies. Using alternative data to estimate the revenues and profits of private companies may be a lucrative endeavor, not only due to the larger universe of private companies but also due to the potentially less efficient pricing of private companies. While public companies are priced in a transparent and competitive market, owners of private companies may not readily know the valuation of their company. Private equity investors may be able to invest in a private company at a lower valuation multiple than a similar public company, especially if that private company sells all or part of its equity without a competitive bidding process.

1.4.1 Alternative data vs. Big data

The definition of big data varies by user, but it is clear that data is big if it can't fit on a single desktop computer. Big data may be defined by its volume and its velocity. Investors using tick data, or the second by second pricing of every stock or option in a market, are likely using big data, as are investors who seek to download a substantial portion of the global news or social media flow in real time. However, some users may be able to process alternative data using a structured database on a single computer, especially if the data has already been cleaned and structured by an alternative data vendor. Rather than processing all of the credit card transactions that may predict a single company's revenue, an investor may simply choose to purchase weekly revenue estimates for each stock from a vendor, who has already organized the data by time and ticker.

There are tradeoffs to using alternative data vendors who clean the data and process the signals. Highly processed data may be more expensive and more widely distributed than the unfiltered raw data, which makes it easier to use. Buyers of aggregated data or buy vs. sell signals experience higher data costs, but lower costs to build the team and infrastructure necessary to clean and organize the data and build their own proprietary signals. While it may be expensive to build this team and infrastructure, investors that incur this expense may be able to find more unique trading signals that may maintain their value for longer periods of time. Data that is widely distributed and easier to use may be deployed by a larger number of investors who may more quickly arbitrage away the value those signals provide.

Let's explore some of the categories of alternative data.

1.4.2 Social Media Postings and Natural Language Processing

A key difference between traditional data and alternative data is that traditional data is typically quantitative in nature. Much of the alternative data is available in text form and must be analyzed using methods such as natural language processing (NLP). Analyzing text can be very messy, as the author of each document or online post has different writing skills, writing styles and some documents may be full of misspellings, acronyms, or even emojis. In 2021, analysts reviewing posts on Reddit needed to determine the meaning of emojis such as money bags, rocket ships, diamonds and hands.

Consider both the value and the difficulty of determining public sentiment regarding a specific company by monitoring social media posts. Application programming interfaces (APIs) are available to read massive numbers of postings to social media sites such as Twitter and Facebook. This data can be valuable in determining sentiment of investors regarding specific companies on a very short term basis, such as hours to weeks. However, the volume of this data can be overwhelming and the comments aren't always mapped to companies. The process for sorting this data is to first determine which posts are related to specific companies and then determine whether the content is expressing a positive or negative sentiment relative to each specific company. While some posters on Twitter make ticker tagging easier by including $TSLA or $FB in a post, other social media platforms to not have posting conventions that are this easy to follow. Once attributing a specific post to a specific company, the analyst next needs to determine the sentiment of that post. The most simple way to determine sentiment is to count the number of positive and negative words. However, this method destroys the context, so more complex NLP methods seek to use chains of words to retain the meaning of phrases. Other challenges of NLP are that different industries have different vocabularies that may need to be learned in order to effectively analyze the text.

Evaluating standard news sources is another key usage of NLP. News may be easier to work with than social media data, as it may be more likely to be tagged or linked to a specific company and may be written in a more formal way that avoids misspellings, emojis and acronyms.

NLP methodology can also be used to interpret company filings with regulators and analysis of audio transcripts from company earnings conference calls. Public companies have obligations to file information with national securities regulators. These include structured data in the form of income statements and balance sheets and unstructured textual data, such as the management discussion and analysis section of an annual report. CEOs and CFOs of publicly traded companies also frequently meet with investors and equity analysts in regular conference calls to discuss quarterly results.

Algorithms can be used to transcribe the voice call into text as well as to mine the voice recording for clues regarding the emotion of the speakers. A key goal of NLP is to compare current filings and calls with those archived from past quarters, focusing on the change in the language, emotion and sentiment from one year to the next.

Both private companies and public companies may have filing requirements with governmental regulators. Governments in many countries sponsor agencies that compile complaints from customers and employees as well as safety issues with pharmaceuticals or manufactured goods. Analysts using NLP tools may be able to predict when the complaints are nearing the point of public disclosure or a mandatory recall of a firm's products. While these signals may be specific to a limited number of industries, such as banking, travel or health care, being able to predict when a regulator will announce potential sanctions on a company can lead to profits on short sales or avoided losses on long positions.

1.4.3 Online Activity

Besides the voluminous social media feeds, other Internet-based information can also be useful for informing investment decisions. Investors may be able to profit from analyzing trends in the behavior of the users of search engines. Services such as Google trends can show changes in the frequency of terms being searched on the Internet. Sharp increases in search activity for a company or one of their products may be indicative of future revenue growth.

Web crawling is used to search and archive data from specific web sites across the Internet. For example, investors may be able to profit from noticing an increase in job postings and hiring activity at a specific firm, which may point to future announcements of revenue growth. Other goals of web crawling may be to find new web pages either through a domain name system (DNS) registration service or for web pages that have been posted by companies that aren't yet fully public. Recently, investors purchased stock in Affirm after noticing a new web page with both Amazon and Affirm in the URL. Having this information just hours or days before an announcement of a partnership between Amazon and Affirm allowed investors to profit nearly 50% on a very short-term trade.

Alternative data may also be used in due diligence for investment managers, such as when a pension fund wishes to invest in a hedge fund. By reviewing information on Revelio Labs, an analyst may note the number of employees at a specific fund that have skills and experience in AI and machine learning. In recent years, hedge funds employing a significant number of skilled quantitative analysts have outperformed the hedge fund industry by an average of 2% to 5%.

1.4.4 Consumer Activity

For companies engaged in business-to-consumer (B2C) transactions, there are many interesting ways to predict revenue. Some alternative data providers access geolocation data from satellites, drones or smartphones. By estimating the foot traffic at retail outlets, investors hope to be able to estimate revenue growth at specific locations where a company's goods are sold.

On a broader scale, consumer spending can also be estimated through emails and credit card transactions. While some credit card processing companies sell aggregate transactions data to estimate GDP or income growth in a certain geography, others allow access to the revenues at specific companies or locations. When reading emails that are sold by email providers, investors are searching for invoices that can be used to estimate the revenue of specific companies.

The online advertising industry makes extensive use of cookies, which are files stored on the computer of each user that document the web sites visited by each user. Online advertisers tend to pay lower prices for general advertisements of small purchases, such as soda manufacturers or specific television shows, but will pay much higher prices when the cookies predict a large purchase is imminent, such as a new car or a piece of business equipment.

1.4.5 Weather and Satellite Data

Some investments may have profits that are correlated to changes in the weather. For example, ski resorts may have lower revenue and profits during times of lower snowfall and higher temperatures, while beach resorts have the opposite problem of greater precipitation and lower temperatures. Those who invest in commodity futures may use satellite photos to estimate crop yields, increasing investments in corn or wheat when photos show reduced crop yields through drought or flood, while reducing investments when it is clear that crop production exceeds market estimates. Farmers can increase their productivity through algorithmically guided tractors that use satellite imagery to optimize the amount of water, seed and fertilizer applied to each 0.1 hectares of land.

1.4.6 Government Statistics

Other sources of alternative data include reports and statistics that have been filed with local and national regulatory agencies. These datasets tend to be specific relative to a given industry. For example, analysts of automobile manufacturers may purchase data from the registration of newly sold vehicles, while investors in home building stocks may be interested in the number of building permits issued for new homes.

1.4.7 Data Ethics

Note that these last two areas of online activity and consumer activity may bring legal risk to the vendor or the user of alternative data. In the case of web scraping, users and vendors of alternative data must understand the policies of the web sites from which the data is being sourced. If the web site specifically prevents harvesting of its data, the user or the vendor may have some legal liability. Data from emails, credit cards and geolocation must carefully deal with personally identifiable information (PII). It is the responsibility of the data vendor to make sure that the data has been anonymized before being sold to the user. Users of alternative data should refuse to accept any dataset where it is clear that the data is non anonymous or the data can somehow be recreated.

1.5 GROWTH OF ALTERNATIVE DATA

As discussed in Section 1.2, in the recent past, new forms and sources of data have emerged in abundance. Companies are using such data to make financial decisions, analyse market behavior, forecast cashflow, identify trends and gain competitive advantage. In comparison with the traditional data, alternative data is generated from broader set of novel sources and provides unique insights into investment opportunities. Demand and usage of alternative data are increasing exponentially. Due to the economy of scale, alternative data becomes more accessible for end users. This chapter provides an overview of growth history and potential of alternative data market, growth drivers and their positive impact.

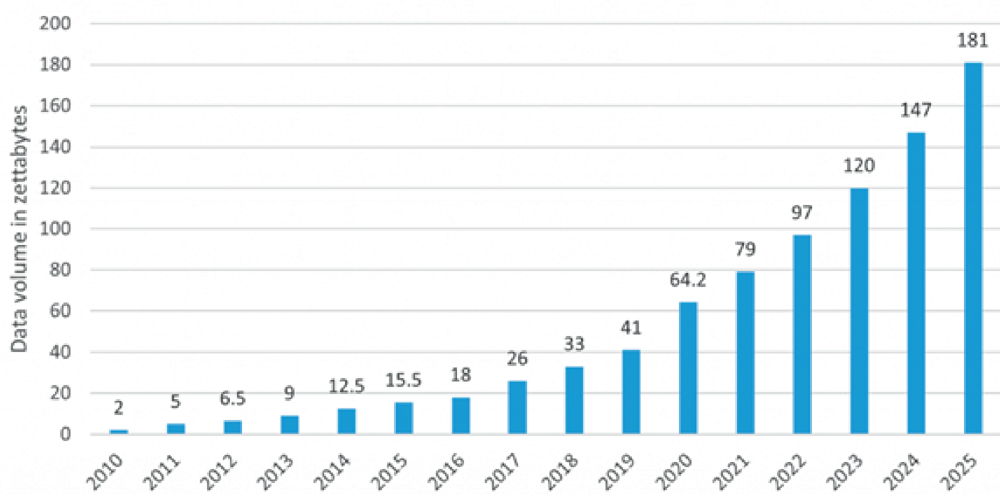


Figure 1.1 Volume of data created and replicated worldwide *(source: IDC)*

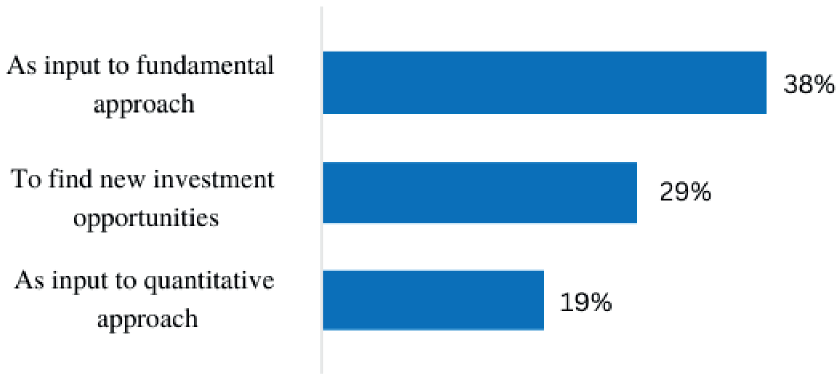


Figure 1.2 Private Equity uses of Alternative Data

In financial market use of alternative data has grown substantially. Applications range from predictive analysis of stock market direction, frequently combined with sentiment analysis, to identification and prevention of financial fraudulent activities. These multiple applications have increased dependence on alternative data. According to survey by Oxylabs and Censuswide, "63% of respondents have started using alternative data to improve their decision making". According to Statista, buy-side companies spent 1.088 billion US dollars on alternative data in 2019. In 2020, 1.708 billion US dollar has been spent. The survey report generated by Global Alternative Data Market in 2020, "the global alternative data market size will grow at a 44% compound annual growth rate (CAGR). By 2026, it will have reached $11.1 billion." According to the report by Grand View Research, in 2020 Alternative Data Market was expected to grow at CAGR of 58.5%, reaching $69.36 billion by 2028.

According to CISCO, 5.3 billion people will use internet services and 29.5 billion networked devices will exist on this planet by 2022. There will be 3.6 global devices and connections per capita [Cisco, 2021]. This increasing number of network devices and internet users clearly predicts the increase in growth of alternative data volume, impacting financial market. According to TabbForum, 38% of data is used as input to fundamental analysis, 29% of it is utilized to find new investment opportunities and 19% is used as input to quantitative approach.

A report published by F. Norrestad at Statista on September 20, 2021 suggested that approximately 50% of hedge funds managers known as Alternative data market leaders utilized more than seven Alternative datasets globally whereas only 8% rest of the market used at least seven alternative datasets [Norrestad, 2020]. This report highlighted the difference between these two classified groups. It further stated: "Using two or more alternative datasets was the most popular approach across both groups with 85% of market leaders and 77% of the rest of the market doing this".

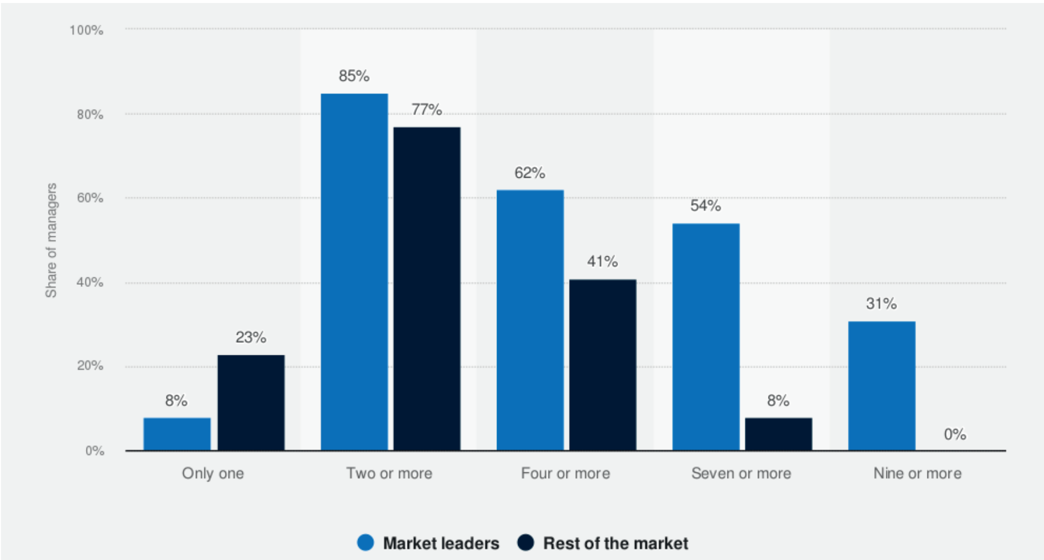


Figure 1.3 Number of alternative datasets used by hedge fund managers worldwide in 2020, by alternative data experience level

Deloitte Center for financial services stated: "Those firms that do not update their investment processes within that time frame could face strategic risks and might very well be outmaneuvered by competitors that effectively incorporate alternative data into their securities valuation and trading signal processes". In today's era, most of the firms rely on Alternative data besides traditional datasets to make their financial decisions more effective.

ALTERNATIVE DATA GROWTH DRIVERS

Increased Number of Smart Cities

Alternative market is expected to grow because of rapid development of smart cities across the globe. Many developed countries are establishing smart cities to manage a city in a systematic manner via data acquisition and analysis. Various alternative data governance schemes are adopted to make the digital life more successful. Smart cities projects are more focusing on empowering cities and developing technologically advanced systems build in this datafied world [Mercille, 2021]. Alternative data market is expected to grow in investment in smart cities. According to Statista, Technology spending on smart city initiatives worldwide will reach 189.5 billion US dollars in 2023.

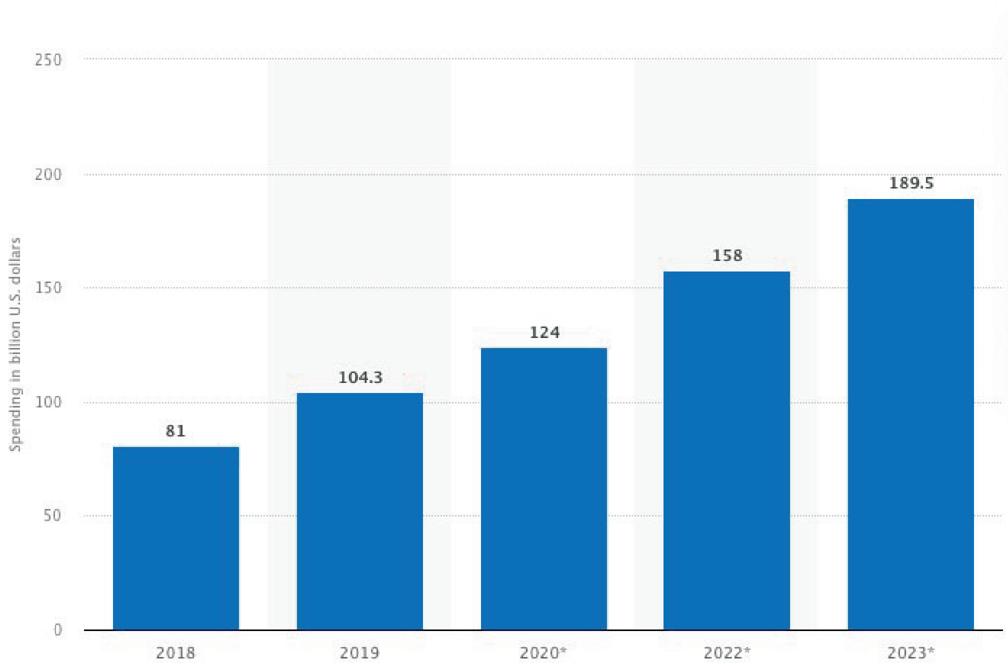


Figure 1.4 Technology spending on smart city initiatives worldwide from 2018 to 2023 (in billion U.S. dollars)

Increase in Internet Penetration

Internet technology and digitalization is increasing day by day; hence, companies are investing more and more in alternative data. The rapid growth of internet penetration paved the path to numerous applications and increases the data acquired from its users. This will ultimately boost the demands of alternative data in the near future.

Increasing Adoption of 5G Networks

5G internet network is one of the hot topics these days, and people are demanding faster internet services. This increasing adoption of 5G network will serve as a driving force for growth of alternative data demands. Better network technologies will accelerate the adoption of alternative data by making high speed data communication and efficient networking. 5G internet technology offers increased bandwidth which will allow low latency rate and more data transmission in a short period of time. According to Statista annual report and prediction, 5G subscription will reach 3 billion by 2025.

Favorable Government Initiatives

Alternative data market growth is significantly impacted by government initiatives to promote technological advancements in various countries across the

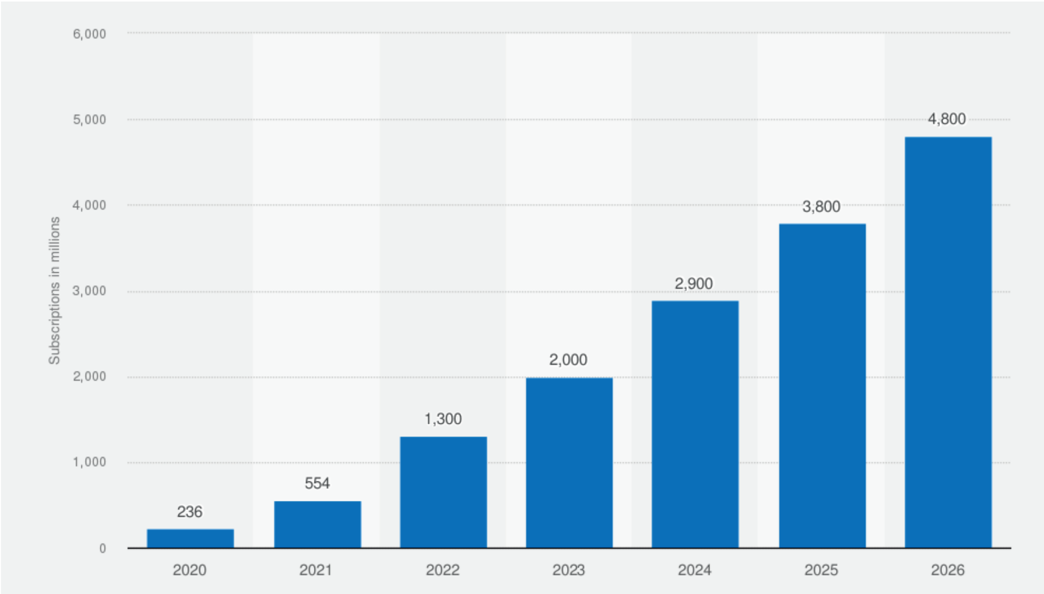


Figure 1.5 Forecast number of mobile 5G subscriptions worldwide from 2019 to 2026

globe. Such projects operate with collaboration with tech firms that focus on use of internet, data security, protection of personal data, risk aversion of data privacy violation and advancement in data sciences. Such initiatives are expected to prove as beneficial impact on alternative data market growth.

Growing Interest in Stock Trading

Alternative data demand is increasing on a greater pace due to rising interest in stock trading. Nowadays, numerous stock trading companies are adopting the use of alternative data approaches for better prediction and outcomes. These novel interests have positive impact on non-traditional and unconventional data acquisition and strategies. These strategies include various floating rate bond funds and bottom-up corporate bonds.

In Conclusion, alternative data is one of the growing domains for research. The use of alternative data in addition to traditional data is used increasingly to make decisions; examples are business expansion, product launch, investment and so on. The major drivers of alternative data include increased number of smart cities, growing interest in stock trading, favorable government initiatives, increased adoption of 5G networks and rising trend of internet penetration. In contrast to conventional data, alternative data is created by means of different sources and gives exceptional knowledge of market trends, customer demands and financial information from different perspectives. More and more companies are using alternative data to make better financial decisions.

1.6 ALTERNATIVE DATA PROVIDERS

As discussed, alternative data is various, unstructured and is extremely large in volume. Despite big potential values that people can acquire by exploiting alternative data, using it is impossible for most individual users or business users who allocate little resource for it. The reason is that the volume, velocity, value, veracity and variety of alternative data make it unavailable to model using, so it needs a lot of effort to be collected and structured. This is the place where alternative data vendors come in to play.

Alternative data vendors or providers are those who solve the above problems and monetize from it. They create the data from their operation or collect data from several sources, then implement a bunch of preparatory organization work to transform the raw data into a form that is ready to be analyzed and pulled insights from.

Like many other industries, alternative data vendors range significantly in size and what they do. At the time this book is written, there are three main kinds of alternative data providers in the market. Firstly, they can be well-known existing market data companies such as Bloomberg. They sell their own alternative datasets and simultaneously create data markets that connect their customers to third-party data vendors. The second and also the most common kind of alternative data vendor is start-ups. They are usually established with a core technical team that collects data from many sources, cleans and transforms data into datasets that can be sold directly to customers or transferred to the customers' system via API technology. Some of them even aggregate many datasets together to construct a trading strategy or produce signals and sell them to investing and trading funds. The last kind of alternative data vendor is big companies which are not traditionally associated with this area but possess exhaust data. Data exhaust refers to the data generated as trails or information byproducts created from all digital or online activities, such as storable choices, actions and preferences such as log files, cookies, temporary files and even information that is generated for every process or transaction done digitally. This data can be very revealing about an individual, so it is very valuable to researchers, marketers and business entities. These companies therefore may sell the data directly to data users or cooperate with other alternative data vendors or a consultancy to help them. One example of this kind of data vendor is Mastercard, which sells its customer transaction data.

The data vendors use their expertise in alternative data processing to monetize these datasets, which includes structuring the data, creating data products, marketing and selling data to users and so on.

The world of alternative data is extremely fragmented, so it is hard for one company to provide all kinds of data. Usually, each data vendor specializes in one or some certain types of alternative data. Below we present some renowned data vendors and their specialties in the market.

Bloomberg is the global leader in business and financial data, news and insight [Bloomberg, 2022]. Using the power of technology, they connect the world's decision makers to accurate information on the financial markets—and help them make faster, smarter decisions.

RavenPack is one of the leading data analytics providers in financial services [RavenPack, 2022]. They help organizations extract value and insights from large amounts of information such as internal documents, emails, instant messages, support tickets and other textual data. Using proprietary NLP technology, RavenPack solves the problem of unstructured and hard-to-process textual data and helps users to extract valuable insights quickly and easily by transforming their unstructured text into structured data. Their process is astoundingly quick and easy.

Yewno deals with creating intelligence data that deals with a knowledge graph [Yewno, 2022]. It is currently being applied across financial services, education and government to help deliver products that make better business decisions. Yewno primarily offers company, research and pricing data.

Tickdata's core product is clean, research—ready, global historical intraday data [TickData, 2022]. Their offering includes institutional-grade quote and trade history from the world's top financial markets, from the Americas to Europe to Africa to Asia to Australia. They cover Equities, Futures, Options, Forex and Cash Indices.

Stocktwits is a large social network for investors and traders, where millions of investors and traders are saying in real time about the stocks, crypto, futures and forex market [StockTwits, 2008]. From this activity, they offer huge tweet databases that can be transformed into sentiment score of these assets, which are then used to create trading rules, perform simulation or back testing, create neural network prediction models and so on.

MarketPsych produces the global standard in financial sentiment and ESG data derived from thousands of news and social media outlets [MarketPsych, 2022]. They develop alpha generation and risk management models based on behavioral principles and also integrate personality testing and practice management tools to help advisors better serve their clients.

Refinitiv is an American-British global provider of financial market data and infrastructure [Refinitiv, 2018]. They provide financial software and risk solutions—delivering news, information and analytics, enabling transactions and connecting the global users. Refinitiv has comprehensive and trusted data covering investment banking, funds, bonds, earnings, macroeconomic indicators, FX, commodities and more, gathering from historical to real-time insights. Their data is delivered via a portfolio of market data products and services, so all users—from traders and investors to wealth and asset managers, as well as risk, compliance, strategy and advisory managers—get the content they need, in their required format.

Alternative data vendors often offer the products as aggregated datasets or as a straight data feed, through APIs. Aggregated data, the less expensive option, is structured, and, therefore, is easier to work with and slot directly into an investment model. But these advantages also make these datasets more widespread to the market, which at the same time is the culprit of their less alpha potential. When there are more people make decision of long or short assets in the market based on the same information, the assets' value will quickly reflex that information, so the alpha added by the information is diminished more quickly.

According to Gene Ekster, CEO of Alternative Data Group and an alternative-data professor at New York University [Gossett, 2021], the aggregated datasets usually lack depth since users will lose the "ability to really dig and mine the data in unique ways". They may also be not truly representative, so users could be swayed by selection bias. Most of the data providers' techniques and methodologies are black-box systems, which are also unavailable for customer inspection, thereby exacerbating aggregate errors due to lack of transparency.

Several years ago, some sportswear retailers inserted an asterisk between the two Lus in their reports: Lu * lulemon, instead of Lululemon. The analyzers didn't have Lu*lu keyword, which made it seem like sales volume had dropped significantly [Gossett, 2021]. This aggregating conclusion led to some short bets, which incurred calamitous loss because Lululemon indeed reported a great quarter. If these people had the raw data, they should have seen the fact and not traded against that.

For these reasons, a raw data feed is considered much more beneficial than aggregated one. But a purely unaltered dataset, with no transformation applied, is essentially just data exhaust. "Any hopes it would provide value would have to be weighed against the considerably heavy clean-up lift" [Gossett, 2021].

The best solution is a live API feed with automatic conversion and structure as much as possible. Under this approach, entity mapping and ticker tagging is a big challenge. This means we have to assign a company reference or brand alias back to its unique stock symbol and correct name. For example, "Facebook" needs to map back to FB and Meta Platforms Inc; "Amazon" needs to map back to AMZN and Amazon.com Inc. And not all references are so direct. Maybe a Twitter user sarcastically references Facebook's or Amazon's slogan while including a typo—"that's powerfull". A hedge fund might want that sentiment included in its investment analysis, but it would need highly sophisticated AI to detect the reference merely. In many cases, it doesn't stop at ticker symbols.

Some alternative data vendors have been focusing on tackling the tagging and mapping challenge. They call their product referential data, which contains all different ways if referencing a given entity, company, or security, mapped back in a way that facilitates the data analysis.

The appearance of alternative data vendors has significantly eased the process of applying alternative data of financial practitioners and open a new genre for alternative data in finance. Financial analysts hinge on alternative data to better their predictions on stocks return. Quality predictions can give an investor a more significant edge over the competition and drive higher profits. Fund managers use alternative data to create signals about the market and entities they are investing on, which help them to enhance investment profit. Hedge funds and investment banks also include alternative data in their risk management process. Some assign ESG values for their assets based on alternative data, from which they can understand fundamental values and assess related risks more precisely. Some use alternative data to produce signals that warn them of potential collapse situations.

Alternative data vendors need to explore and build datasets that are valuable to and recognized among their target customers, so that their data can be monetized.

However, there are differences in the quality and price of data. Some data intermediaries even go for a quantity-over-quality approach while aggregating datasets with high ticker coverage but not necessarily insightful ticker coverage. Therefore, not every data is valuable to users. One might put a lot of effort and resources into exploiting and analyzing a dataset but gain no value from it, or the profit it creates cannot cover the huge cost of utilizing it. Any ability to diminish the turnaround time between acquisition and analysis is valuable.

People start using alternative data may keep questioning how they know if a dataset is going to be valuable. It could take six months of R&D, and they have to buy it first. They will not know how much additional profit the data is going to create until much later.

Neuravest is one of the companies focused on cracking that enigma and may be considered an intermediary of the intermediaries. They associate with 42 other alternative-data vendors and work to validate datasets before integrating them into machine-learning investment models.

Raw data is fed into the system to generate a data qualification report as what the company calls it. In this system, the data is measured along 12 checkpoints before being reasonably qualified to be incorporated into a model. These 12 checkpoints include a time indicator which measure the time length before a signal drops value, a price action distribution following a given event, such as a news announcement that stimulates social-media conversation.

After validation, the data is cleaned, tagged and normalized before building the model to create testable investment arguments. By aggregating uncorrelated data sets, the models aim to identify constituent stocks and assets that are about to experience unusual volatility relative to similar stocks.

With the growing market size for alternative data, more and more players are entering, providing alternative data for investors and hedge fund managers. As we are living in an extremely competitive buying and selling environment, it is important for everyone to capture whatever data they can get their hands on to gain a competitive advantage.

1.7 CHALLENGES OF USING ALTERNATIVE DATA

Users of traditional, simpler and structured data usually end up building models with correlated signals and investing in trades with eroding alpha. The more users of a dataset and the more structured the datasets are, the harder it is to get a return advantage because so many other investors have access to the same data. A crowded trade also has risk management implications on exit. The quest for differentiated signals and a longer-lasting return advantage is making contemporary investors gravitate towards alternative data. The potential rewards of deploying alternative data are often accompanied by challenges, many of which can be significant.

The sheer volume and richness of alternative data mean that investment managers have a likely opportunity to come up with unique signals and potentially more sustainable return advantages. However, the challenges of using alternative data are also directly related to the volume and complexity of the data. Noise and missing values also stymie their facile, instant use. It takes substantial effort to clean the data, map the data to specific securities and test the data within a firm or portfolio's existing frameworks (sets of data and models). Unfortunately, not every dataset will be additive to the models currently employed; so, factor in wasted effort to clean, load, compare and test datasets that may never be deployed. Also, users of alternative data will need to perform a cost-benefit analysis before employing the data. The costs are not only to purchase the data, but also to build a team to test, clean and load the data. The benefits are not the performance of the data on a stand-alone basis, but the value when added to an investor's existing set of frameworks. As more models, factors, and data are included in a trading system, the more challenging it is to find the next framework that is value-additive. Generating innovative insights by mashing up multiple datasets is one of the key challenges for any alt data expert, yet the results can add tremendous amount of value to their investment debate.

Some examples come to mind. On a Monday in late April 2019, a boutique research firm—using publicly available corporate aircraft and flight data—reported that a Gulfstream V jet belonging to Occidental Petroleum had been sighted the day before in Omaha, Nebraska, the hometown of famed investor Warren Buffett. A day later, Buffett's Berkshire Hathaway announced its intention to invest $10 billion in a preferred stake in Houston-based Occidental, a move that ultimately proved critical in helping Occidental trump Chevron in a takeover battle for Anadarko Petroleum. More immediately, however, Berkshire's official announcement of its involvement amplified price movements and trading volume in Occidental and Chevron shares relative to the prior day—when information about the jet data was known to just a handful of eagle-eyed investment sleuths.

Aircraft and flight logs are examples of "alternative data"—non-reportable information that some investors increasingly are embracing to gain an edge. Other examples include private opinion polls before a key political referendum; satellite and drone imagery of ports and mall parking lots; credit-card transactions; job postings; patent and trademark filings; and social-media sentiment, including ratings and reviews of products and services. Unlike financial and other data reportable to market regulators, which is readily available and widely disseminated, alternative data is often proprietary, affording its collectors or buyers information that other investors do not have.

Yet, the leap from today's alternative data to a future investment consensus is neither immediate nor guaranteed. Data vintage, authority and provenance matter. Any changes in the data generation or collection process also can cause problems for those relying on the information to make investment decisions.

Just ask Bloomberg, which constructed and published a Tesla Model 3 production volume estimator, based on Vehicle Identification Numbers (VINs) from official U.S. government sources, social media reports and input from owners of Tesla's electric cars. The Model 3 Tracker launched in February 2018. For most of 2018, as the market focused on Tesla's production ramp-up for the Model 3, Bloomberg's alternative-data-based estimate proved a good predictor of the company's quarterly output. Model 3 production numbers were a key driver of Tesla's stock price in late 2018 as well as in 2019's first quarter.

In early 2019, however, a rise in Tesla exports led to changes in how many extra VINs the company registered pre-production to keep on hand a buffer of the numbers. These data—generation and collection—changes caused the Bloomberg tracker to veer off course, overestimating production numbers for the first quarter of 2019 by 26%. Tesla stock sold off early April of that year, when the actual production and delivery numbers were announced, catching many investors relying on the Bloomberg tracker off guard.

To cite another example, the 2016 U.S. Presidential Election was a significant black eye for election forecasters. Relying on alternative data, generated via opinion polls, almost all the pundits had predicted Hillary Clinton to be the victor. Similarly, the Brexit vote during June of 2016 and the election of Scott Morrison as the Australian Prime Minister were additional instances where alternative data-informed methods led analysts astray. Investors across many asset classes—ranging from currencies to equities—were stunned by the volatility resulting from the surprise election results. Many firms and pundits have recently talked about backing off from election forecasting.

New strategic surprises also expose new data streams. As the recent pandemic unleashed its fury, new sources of data (e.g., the number of people going through the airport TSA checkpoints daily in the US) emerged. Investors started to focus on the robustness of the reopening across different sectors of the economy and were looking for multiple, new clues.

AI has recently emerged as a premium refiner of alternative data, especially to deal with surprises arising from missing, incomplete and uncertain values; and, to normalize for transient effects, such as traffic-flow numbers affected by inclement weather, health warnings or construction.

However, maximizing the utility of alternative data likely requires a hybrid analysis, combining humans and machines. Computers are able to process large amounts of data rapidly, while humans are able to apply judgment and empathy in ambiguous situations, often guided by past experiences, some from unrelated domains. Just as a pilot must interpret conflicting readings or flight-deck alerts in the cockpit, an analyst should use alternative data in concert with other information to make better choices or mitigate risk. Fund managers making significant use of AI and alternative data have shown the potential to increase returns relative to managers who are not yet making use of these tools. They have done so, by typically using more unique and often unstructured datasets, and by employing AI to scale human expertise. Prudent use of AI can help overcome human fatigue and short attention spans, helping an investor build a lasting advantage using carefully curated alternative data.

There are a few significant caveats to keep in mind while employing alternative data. The first is that the data needs to imply a relatively direct connection to asset prices—a potential takeover usually implies a higher price for the company being acquired, an uptick in implied product demand points to increasing revenues for the manufacturing company in the near term; and, a candidate's improved election forecast may point to imminent implementation of previously telegraphed policy changes. The second caveat is that often there is a significant delay between what is hinted by the data and a majority of the market participants acting on or realizing it. Consensus view or stock sentiment may not change overnight, and an asset can continue to remain mispriced. If the adjustment to pricing of a security takes a long time to materialize, it erodes the internal rate of return (IRR), turning a potentially good investment decision into a mediocre one.

Finally, as previously stated, there is also the challenge of dealing with the data in an ethical manner. Ethical harvesting of data is performed when all of the data is sourced legally from sources that do not prohibit the use of the data. Users of alternative data are expected to comply with Global Data Protection Regulation (GDPR) and other rules regarding data usage and storage. It is imperative that users of alternative data work with completely anonymized data and work with vendors who are compliant with GDPR and other regulations; and are highly respectful of protecting PIII.

Caveat (alternative data) emptor!

1.8 SUMMARY DISCUSSION AND CONCLUSION

A few decades ago, a financial modeler and an analyst depended mainly on small number of traditional data sources to analyze and make decisions. In recent times, this fast-evolving field of alternative data as used in financial analytics is emerging as a major disruptor of traditional applications of financial modelling. Thus all instances such as decision-making in asset allocation, predicting the market, risk management and credit scoring are one way or another affected. In this Handbook we have focused mainly on the above areas of financial applications. We have highlighted that many financial (expert) practitioners have already used alternative data to create trading strategies, investment strategies and risk management applications. We believe this field will grow rapidly and contents of this Handbook provides the readers an early entry into this exciting field.

REFERENCES

Aldridge, I. (2015). Trends: all finance will soon be big data finance.

Barigozzi, M. and Hallin, M. (2017). A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):581–605.

Berg, T., Burg, V., Gombović, A., and Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7):2845–2897.

Bertolotti, A. (2020). Effectively managing risks in an esg portfolio. *Journal of Risk Management in Financial Institutions*, 13(3):202–211.

Bloomberg (2022). Bloomberg. https://www.bloomberg.com/. Accessed: 2022-06-23.

Brown, L. D. (1997). Analyst forecasting errors: Additional evidence. *Financial Analysts Journal*, 53(6):81–88.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Cheney, J. S. (2008). Alternative data and its use in credit scoring thin-and no-file consumers. *FRB of Philadelphia-Payment Cards Center Discussion Paper* (08-01).

Cisco, U. (2021). Cisco annual internet report (2018–2023) white paper. 2020. *Acessado em*, 10(01).

Cong, L. W., Li, B., and Zhang, Q. T. (2021). Alternative data in fintech and business intelligence. In *The Palgrave Handbook of FinTech and Blockchain*, pages 217–242. Springer.

De Spiegeleer, J., Höcht, S., Jakubowski, D., Reyners, S., and Schoutens, W. (2021). Esg: a new dimension in portfolio allocation. *Journal of Sustainable Finance & Investment*, pages 1–41.

Djeundje, V. B., Crook, J., Calabrese, R., and Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1(2000).

Dourish, P. and Gomez Cruz, E. (2018). Datafication and data fiction: narrating data and narrating with data. *Big Data & Society*, 5(2).

Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica*, 68(1):1–22.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2):427–465.

Fan, J., Lv, J., and Qi, L. (2011). Sparse high dimensional models in economics. *Annual Review of Economics*, 3:291.

Fang, B. and Zhang, P. (2016). Big data in finance. In *Big Data Concepts, Theories, and Applications*, pages 391–412. Springer.

Goldfarb, A. and Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1):3–43.

Goldstein, I., Spatt, C. S., and Ye, M. (2021). Big data in finance. *The Review of Financial Studies*, 34(7):3213–3225.

Gossett, S. (2021). What is alternative data and why is it changing finance? *Built In*.

Henriksson, R., Livnat, J., Pfeifer, P., and Stumpp, M. (2019). Integrating esg in portfolio construction. *The Journal of Portfolio Management*, 45(4):67–81.

In, S. Y., Rook, D., and Monk, A. (2019). Integrating alternative data (also known as esg data) in investment decision making. *Global Economic Review*, 48(3):237–260.

Jame, R., Johnston, R., Markov, S., and Wolfe, M. C. (2016). The value of crowdsourced earnings forecasts. *Journal of Accounting Research*, 54(4):1077–1110.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Kolanovic, M. and Krishnamachari, R. T. (2017). Big data and AI strategies: machine learning and alternative data approach to investing. *JP Morgan Global Quantitative & Derivatives Strategy Report*.

Lewis, M. (2014). The wolf hunters of wall street: an adaptation from'flash boys: A wall street revolt,'by michael lewis. *The New York Times Magazine March*, 3.

MarketPsych (2022). Marketpsych: financial markets sentiment and esg data. https://www.marketpsych.com/. Accessed: 2022-06-23.

Mercille, J. (2021). Inclusive smart cities: beyond voluntary corporate data sharing. *Sustainability*, 13(15):8135.

Mitra, G., Baigurmanova, D., and Jhunjhunwala, S. (2020). The new landscape of financial markets: Participants and asset classes. *Available at SSRN 4104690*.

Monk, A., Prins, M., and Rook, D. (2019). Rethinking alternative data in institutional investment. *The Journal of Financial Data Science*, 1(1):14–31.

Norrestad, F. (2020). Alternative data: Hedge fund manager use globally 2020. https://www.statista.com/statistics/1169968/alternative-data-hedge-fund-managers-global/.

O'Leary, D. E. (2013). Big data', the 'internet of things'and the 'internet of signs. *Intelligent Systems in Accounting, Finance and Management*, 20(1):53–65.

Quinlan, B., Kwan, Y., and Cheng, H. (2017). Alternative alpha: unlocking hidden value in the everyday.

RavenPack (2022). Ravenpack: technology and insights for data-driven companies. https://www.ravenpack.com/. Accessed: 2022-06-23.

Refinitiv (2018). Refinitiv financial solutions – financial technology experts. https://www.refinitiv.com/. Accessed: 2022-06-23.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review*, pages 289–315.

Stock, J. H. and Watson, M. (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, 173:205.

StockTwits (2008). Stocktwits – the largest community for investors and traders. https://stocktwits.com/. Accessed: 2022-06-23.

Syed, A., Gillela, K., and Venugopal, C. (2013). The future revolution on big data. *Future*, 2(6):2446–2451.

The Organization for Economic Cooperation and Development (2013). *Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data"*. OECD Digital Economy Papers.

TickData (2022). Tick data. https://www.tickdata.com/. Accessed: 2022-06-23.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208(1):231–248.

Wang, P. (2008). *Financial Econometrics*. Routledge.

Xiao, C. and Chen, W. (2018). Trading the Twitter sentiment with reinforcement learning. *arXiv preprint arXiv:1801.02243*.

Yewno (2022). Yewno – transforming information into knowledge. https://www.yewno.com/. Accessed: 2022-06-23.

Alternative Data: Overview

Aldridge, I. (2015). Trends: all finance will soon be big data finance.

Barigozzi, M. and Hallin, M. (2017). A network analysis of the volatility of high dimensional financial series. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66(3):581–605.

Berg, T. , Burg, V. , Gombović, A. , and Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7):2845–2897.

Bertolotti, A. (2020). Effectively managing risks in an esg portfolio. Journal of Risk Management in Financial Institutions, 13(3):202–211.

Bloomberg (2022). Bloomberg. https://www.bloomberg.com/. Accessed: 2022-06-23.

Brown, L. D. (1997). Analyst forecasting errors: Additional evidence. Financial Analysts Journal, 53(6):81–88.

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of finance, 52(1):57–82.

Cheney, J. S. (2008). Alternative data and its use in credit scoring thin-and no-file consumers. *FRB of Philadelphia-Payment Cards Center Discussion Paper* (08-01).

Cisco, U. (2021). Cisco annual internet report (2018–2023) white paper. 2020. Acessado em, 10(01).

Cong, L. W. , Li, B. , and Zhang, Q. T. (2021). Alternative data in fintech and business intelligence. In The Palgrave Handbook of FinTech and Blockchain, pages 217–242. Springer.

De Spiegeleer, J. , Höcht, S. , Jakubowski, D. , Reyners, S. , and Schoutens, W. (2021). Esg: a new dimension in portfolio allocation. Journal of Sustainable Finance & Investment, pages 1–41.

Djeundje, V. B. , Crook, J. , Calabrese, R. , and Hamid, M. (2021). Enhancing credit scoring with alternative data. Expert Systems with Applications, 163.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 1(2000).

Dourish, P. and Gomez Cruz, E. (2018). Datafication and data fiction: narrating data and narrating with data. Big Data & Society, 5(2).

Engle, R. F. (2000). The econometrics of ultra-high-frequency data. Econometrica, 68(1):1–22.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. The Journal of Finance, 47(2):427–465.

Fan, J. , Lv, J. , and Qi, L. (2011). Sparse high dimensional models in economics. Annual Review of Economics, 3:291.

Fang, B. and Zhang, P. (2016). Big data in finance. In Big Data Concepts, Theories, and Applications, pages 391–412. Springer.

Goldfarb, A. and Tucker, C. (2019). Digital economics. Journal of Economic Literature, 57(1):3–43.

Goldstein, I. , Spatt, C. S. , and Ye, M. (2021). Big data in finance. The Review of Financial Studies, 34(7):3213–3225.

Gossett, S. (2021). What is alternative data and why is it changing finance? *Built In*.

Henriksson, R. , Livnat, J. , Pfeifer, P. , and Stumpp, M. (2019). Integrating esg in portfolio construction. The Journal of Portfolio Management, 45(4):67–81.

In, S. Y. , Rook, D. , and Monk, A. (2019). Integrating alternative data (also known as esg data) in investment decision making. Global Economic Review, 48(3):237–260.

Jame, R. , Johnston, R. , Markov, S. , and Wolfe, M. C. (2016). The value of crowdsourced earnings forecasts. Journal of Accounting Research, 54(4):1077–1110.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. The Journal of Finance, 48(1):65–91.

Kolanovic, M. and Krishnamachari, R. T. (2017). Big data and AI strategies: machine learning and alternative data approach to investing. *JP Morgan Global Quantitative & Derivatives Strategy Report*.

Lewis, M. (2014). The wolf hunters of wall street: an adaptation from'flash boys: A wall street revolt,'by michael lewis. The New York Times Magazine March, 3.

MarketPsych (2022). Marketpsych: financial markets sentiment and esg data. https://www.marketpsych.com/. Accessed: 2022-06-23.

Mercille, J. (2021). Inclusive smart cities: beyond voluntary corporate data sharing. Sustainability, 13(15):8135.

Mitra, G. , Baigurmanova, D. , and Jhunjhunwala, S. (2020). The new landscape of financial markets: Participants and asset classes. *Available at SSRN 4104690*.

Monk, A. , Prins, M. , and Rook, D. (2019). Rethinking alternative data in institutional investment. The Journal of Financial Data Science, 1(1):14–31.

Norrestad, F. (2020). Alternative data: Hedge fund manager use globally 2020. https://www.statista.com/statistics/1169968/alternative-data-hedge-fund-managers-global/.

O'Leary, D. E. (2013). Big data', the ‘internet of things'and the ‘internet of signs. Intelligent Systems in Accounting, Finance and Management, 20(1):53–65.

Quinlan, B. K. Y. and Cheng, H. (2017). Alternative alpha: unlocking hidden value in the everyday.

RavenPack (2022). Ravenpack: technology and insights for data-driven companies. https://www.ravenpack.com/. Accessed: 2022-06-23.

Refinitiv (2018). Refinitiv financial solutions – financial technology experts. https://www.refinitiv.com/. Accessed: 2022-06-23.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? Accounting Review, pages 289–315.

Stock, J. H. and Watson, M. (2009). Forecasting in dynamic factor models subject to structural instability. The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry, 173:205.

StockTwits (2008). Stocktwits – the largest community for investors and traders. https://stocktwits.com/. Accessed: 2022-06-23.

Syed, A. , Gillela, K. , and Venugopal, C. (2013). The future revolution on big data. Future, 2(6):2446–2451.

The Organization for Economic Cooperation and Development (2013). *Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by “Big Data”*. OECD Digital Economy Papers.

TickData (2022). Tick data. https://www.tickdata.com/. Accessed: 2022-06-23.

Wang, D. , Liu, X. , and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. Journal of Econometrics, 208(1):231–248.

Wang, P. (2008). Financial Econometrics. Routledge.

Xiao, C. and Chen, W. (2018). Trading the Twitter sentiment with reinforcement learning. *arXiv preprint arXiv:1801.02243*.

Yewno (2022). Yewno – transforming information into knowledge. https://www.yewno.com/. Accessed: 2022-06-23.

Contemplation and Reflection on Using Alternative Data for Trading and Fund Management

Authority, U. S. (2021). UK Energy in Brief. http://www.gov.uk/government/statistics/uk-energy-in-brief-2021.
Babii, A. , Ghysels, E. , and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. Journal of Business & Economic Statistics, 40(3):1094–1106.
Bailey, D. H. , Borwein, J. M. , de Prado, M. L. , and Zhu, Q. J. (2014). Pseudomathematics and financial charlatanism: the effects of backtest over fitting on out-of-sample performance. Notices of the AMS, 61(5):458–471.
Banbura, M. , Giannone, D. , Modugno, M. , and Reichlin, L. (2013). Now-casting and the real-time data flow. In Handbook of Economic Forecasting, volume 2, pages 195–237. Elsevier.
Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10):78–87.
Harvey, C. R. (2017). Presidential address: the scientific outlook in financial economics. The Journal of Finance, 72(4):1399–1440.
Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. *Advances in Neural Information Processing Systems* , 30:4768–4777.
Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 160(1):246–256.
Ribeiro, M. T. , Singh, S. , and Guestrin, C. (2016). “why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* , pages 1135–1144.
Romano, J. P. , Shaikh, A. M. , Wolf, M. , et al. (2010). Multiple testing. *The New Palgrave Dictionary of Economics (Online Edition).*

Global Economy and Markets Sentiment Model

Casanova, C. , Ortiz, A. , Rodrigo, T. , Xia, L. , Iglesias, J. , et al. (2017). Tracking Chinese vulnerability in real time using Big Data. https://www.bbvaresearch.com/wp-content/uploads/2017/03/20170328_CVSI2.pdf.
Diakoulaki, D. , Mavrotas, G. , and Papayannakis, L. (1995). Determining objective weights in multiple criteria problems: the critic method. Computers & Operations Research, 22(7):763–770.
Peterson, R. L. (2016). Trading on Sentiment: The Power of Minds Over Markets. John Wiley & Sons.
RavenPack (2021). Developed markets sovereign bonds investing-enhancing style with sentiment. Wilmott, 2021(116):10–13.

Enhanced Corporate Bond Yield Modeling Incorporating Macroeconomic News Sentiment

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723.
Apergis, N. (2015). Forecasting credit default swaps (cdss) spreads with newswire messages: evidence from European countries under financial distress. Economics Letters, 136:92–94.
Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. The Journal of Finance, 61(4):1645–1680.
Bessembinder, H. , Spatt, C. , and Venkataraman, K. (2020). A survey of the microstructure of fixed-income markets. Journal of Financial and Quantitative Analysis, 55(1):1–45.
Box, G. E. and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control, *revised ed.* Holden-Day.
diBartolomeo, D. (2016). Credit risk assessment of corporate debt using sentiment and news. In Mitra, G. and Yu, X. , editors, Handbook of Sentiment Analysis in Finance. Optirisk Systems in collaboration with Albury Books.
Fernandes, C. , Gama, P. M. , and Vieira, E. (2016). Does local and euro area sentiment matter for sovereign debt markets? Evidence from a bailout country. Applied Economics, 48(9):816–834.
Fuller, W. A. (1976). Introduction to Statistical Time Series. John Wiley & Sons, New York.
Laborda, R. and Olmo, J. (2014). Investor sentiment and bond risk premia. Journal of Financial Markets, 18:206–233.
MarketAxess (2016). 4q 2015 fact sheet.
Nagel, J. (2016). Markets committee: electronic trading in fixed income markets. Technical report, Working paper, Tech. rep. 9789291974207, Bank for International Settlements.
Nayak, S. (2010). Investor sentiment and corporate bond yield spreads. Review of Behavioral Finance, 2(2):59–80.
O'Hara, M. and Zhou, X. A. (2021). The electronic evolution of corporate bond dealers. Journal of Financial Economics, 140(2):368–390.
Svensson, L. E. (1994). Estimating and Interpreting Forward Interest Rates: Sweden 1992–1994 (No. w4871). National Bureau of Economic Research.
Yu, X. (2014). Analysis of News Sentiment and Its Applications to Finance. PhD thesis, Brunel University London, UK.
Yu, X. and Mitra, G. (2016). An impact measure for news: its use in (daily) trading strategies. In The Handbook of Sentiment Analysis in Finance, pages 288–309. Albury Books.

Artificial Intelligence, Machine Learning and Quantitative Models

Abraham, R. , Samad, M. E. , Bakhach, A. M. , El-Chaarani, H. , Sardouk, A. , Nemar, S. E. , and Jaber, D. (2022). Forecasting a stock trend using genetic algorithm and random forest. Journal of Risk and Financial Management, 15(5):188.

Alberg, D. , Shalit, H. , and Yosef, R. (2008). Estimating stock market volatility using asymmetric garch models. Applied Financial Economics, 18(15):1201–1208.

Ali, S. S. , Mubeen, M. , Lal, I. , and Hussain, A. (2018). Prediction of stock performance by using logistic regression model: evidence from pakistan stock exchange (psx). Asian Journal of Empirical Research, 8(7):247–258.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural computation, 9(7):1545–1588.

Ampomah, E. K. , Qin, Z. , and Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information, 11(6):332.

Anand, C. (2021). Comparison of stock price prediction models using pre-trained neural networks. Journal of Ubiquitous Computing and Communication Technologies (UCCT), 3(02):122–134.

Bell, D. E. , Raiffa, H. , and Tversky, A. (1988). Decision Making: Descriptive, Normative, and Prescriptive Interactions. Cambridge University Press.

Bishop, C. M. and Nasrabadi, N. M. (2006). Pattern Recognition and Machine Learning, volume 4. Springer.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3):307–327.

Booth, A. , Gerding, E. , and McGroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. Expert Systems with Applications, 41(8):3651–3661.

Bouras, C. , Christou, C. , Gupta, R. , and Suleman, T. (2019). Geopolitical risks, returns, and volatility in emerging stock markets: evidence from a panel garch model. Emerging Markets Finance and Trade, 55(8):1841–1856.

Brahme, A. (2014). Comprehensive biomedical physics. Newnes.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2):123–140.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Brooks, C. (2014a). In C brooks. Introductory Econometrics for Finance, 331.

Brooks, C. (2014b). Introductory Econometrics for Finance. 3rd *Edition*. Cambridge University Press.

CBOE (2016). The VIX Index. https://www.cboe.com/tradable_products/vix/faqs/, Last accessed on 05 December 2022.

Chen, W. , Jiang, M. , Zhang, W.-G. , and Chen, Z. (2021). A novel graph convolutional feature based convolutional neural network for stock trend prediction. Information Sciences, 556:67–94.

Chiang, T. C. and Doong, S.-C. (2001). Empirical analysis of stock returns and volatility: Evidence from seven asian stock markets based on tar-garch model. Review of Quantitative Finance and Accounting, 17(3):301–318.

Choudhry, T. (2000). Day of the week effect in emerging asian stock markets: evidence from the garch model. Applied Financial Economics, 10(3):235–242.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica: Journal of the Econometric Society, pages 987–1007.

Fama, E. F. (1995). Random walks in stock market prices. Financial Analysts Journal, 51(1):75–80.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. Information and computation, 121(2):256–285.

Gong, J. and Sun, S. (2009). A new approach of stock price prediction based on logistic regression model. In *2009 International Conference on New Trends in Information and Service Science* , pages 1366–1371. IEEE.

Gyamerah, S. A. , Ngare, P. , and Ikpe, D. (2019). On stock market movement prediction via stacking ensemble learning method. In *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)* , pages 1–8. IEEE.

Hastie, T. , Tibshirani, R. , Friedman, J. H. , and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2.

Ho, T. K. (1995). -block Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* , volume 1, pages 278–282. IEEE.

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8):832–844.

Hu, H. , Tang, L. , Zhang, S. , and Wang, H. (2018). Predicting the direction of stock markets using optimized neural networks with google trends. Neurocomputing, 285:188–195.

Hull, J. C. (2003). Options futures and other derivatives. Pearson Education India.

Hung, J.-C. (2011). Applying a combined fuzzy systems and garch model to adaptively forecast stock market volatility. Applied Soft Computing, 11(5):3938–3945.

Jain, A. K. , Mao, J. , and Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. Computer, 29(3):31–44.

James, G. , Witten, D. , Hastie, T. , and Tibshirani, R. (2013). An Introduction to Statistical Learning, volume 112. Springer.

Jiang, M. , Liu, J. , Zhang, L. , and Liu, C. (2020). An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. Physica A: Statistical Mechanics and its Applications, 541:122272.

Kahneman, D. and Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In Handbook of the Fundamentals of Financial Decision Making: Part I, pages 99–127. World Scientific.

Kahnemann, D. (1979). Prospect theory: an analysis of decision under risk. Econometrica, 47:263–292.

Kara, Y. , Boyacioglu, M. A. , and Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the istanbul stock exchange. Expert systems with Applications, 38(5):5311–5319.

Lee, S. (2004). Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using gis. Environmental Management, 34(2):223–232.

Lin, W. , Wu, Z. , Lin, L. , Wen, A. , and Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. IEEE Access, 5:16568–16575.

Lintner, J. (1969). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets: a reply. The Review of Economics and Statistics, pages 222–224.

Lohrmann, C. and Luukka, P. (2019). Classification of intraday s&p500 returns with a random forest. International Journal of Forecasting, 35(1):390–407.

Luo, S.-S. , Weng, Y. , Wang, W.-W. , and Hong, W.-X. (2017). L1-regularized logistic regression for event-driven stock market prediction. In *2017 12th International Conference on Computer Science and Education (ICCSE)* , pages 536–541. IEEE.
Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1):77–91.
Miller, M. H. and Modigliani, F. (1961). Dividend policy, growth, and the valuation of shares. The Journal of Business, 34(4):411–433.
Mitra, G. , Greenberg, H. J. , Lootsma, F. A. , Rijkaert, M. J. , and Zimmermann, H. J. (1988). Mathematical Models for Decision Support. Springer.
Mitra, G. and Mitra, L. (2011). The Handbook of News Analytics in Finance. John Wiley & Sons, Chichester, UK.
Mitra, G. and Yu, X. (2016a). The Handbook of Sentiment Analysis in Finance. Optirisk Systems in collaboration with Albury Books.
Mitra, G. and Yu, X. (2016). Handbook of Sentiment Analysis in Finance. Optirisk Systems in collaboration with Albury Books.
Neumann, J. V. and Morgenstern, O. (1944). Theory of games and economic behavior Princeton: Princeton Univ. Press.
Nti, I. K. , Adekoya, A. F. , and Weyori, B. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. Journal of Big Data, 7(1):1–40.
Nti, K. O. , Adekoya, A. , and Weyori, B. (2019). Random forest based feature selection of macroeconomic variables for stock market prediction. American Journal of Applied Sciences, 16(7):200–212.
Priddy, K. L. and Keller, P. E. (2005). Artificial Neural Networks: An Introduction, volume 68. SPIE press.
Qin, L. and Ren, R. (2019). Integrating market sentiment with trading rules-empirical study on china stock market. In *2019 International Conference on Pedagogy, Communication and Sociology (ICPCS 2019)* , pages 358–363. Atlantis Press.
Rakhra, M. , Soniya, P. , Tanwar, D. , Singh, P. , Bordoloi, D. , Agarwal, P. , Takkar, S. , Jairath, K. , and Verma, N. (2021). Crop price prediction using random forest and decision tree regression: a review. *Materials Today: Proceedings*.
Sadorsky, P. (2021). A random forests approach to predicting clean energy stock prices. Journal of Risk and Financial Management, 14(2):48.
Sagi, O. and Rokach, L. (2018). Ensemble learning: a survey wiley interdisciplinary reviews: data mining and knowledge discovery. 8(4):e1249.
Scholes, M. and Black, F. (1973). The pricing of options and corporate liabilities. Journal of Political Economy, 81(3):637–654.
SenRisk (2018). Sentiment analysis and risk assessment. http://senrisk.eu/. Last accessed on 05 December 2022.
Serrano, W. (2022). The random neural network in price predictions. Neural Computing and Applications, 34(2):855–873.
Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. The Journal of Finance, 19(3):425–442.
Shastri, M. , Roy, S. , and Mittal, M. (2019). Stock price prediction using artificial neural model: an application of big data. EAI Endorsed Transactions on Scalable Information Systems, 6(20):e1–e1.
Smales, L. A. (2014). News sentiment in the gold futures market. Journal of Banking & Finance, 49:275–286.
Srinu Vasarao, P. and Chakkaravarthy, M. (2022). Time series analysis using random forest for predicting stock variances efficiency. In Intelligent Systems and Sustainable Computing, pages 59–67. Springer.
Stacking (2020). Ensembling technique to improve prediction. https://medium.com/ml-research-lab/stacking-ensemble-meta-algorithms-for-improve-predictions-f4b4cf3b9237. Last accessed on 07 December 2022.
Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press.
Tan, Z. , Yan, Z. , and Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the chinese stock market. Heliyon, 5(8):e02310.
Thakkar, A. and Chaudhari, K. (2020). Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks. Applied Soft Computing, 96:106684.
Thakkar, A. and Chaudhari, K. (2021). A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. Expert Systems with Applications, 177:114800.
Weng, B. , Lu, L. , Wang, X. , Megahed, F. M. , and Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. Expert Systems with Applications, 112:258–273.
Wolpert, D. H. (1992). Stacked generalization. Neural networks, 5(2):241–259.
Xu, L. , Kinkyo, T. , and Hamori, S. (2018). Predicting currency crises: A novel approach combining random forests and wavelet transform. Journal of Risk and Financial Management, 11(4):86.
Xu, Y. , Yang, C. , Peng, S. , and Nojima, Y. (2020). A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning. Applied Intelligence, 50(11):3852–3867.
Yegnanarayana, B. (2009). Artificial Neural Networks. PHI Learning Pvt. Ltd.
Zhang, D. and Lou, S. (2021). The application research of neural network and bp algorithm in stock price pattern classification and prediction. Future Generation Computer Systems, 115:872–879.
Zhou, F. , Zhang, Q. , Sornette, D. , and Jiang, L. (2019). Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. Applied Soft Computing, 84:105–747.
Zi, R. , Jun, Y. , Yicheng, Y. , Fuxiang, M. , and Rongbin, L. (2022). Stock price prediction based on optimized random forest model. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)* , pages 777–783. IEEE.

Asset Allocation Strategies: Enhanced by Micro-Blog

Al Nasseri, A. , Tucker, A. , and de Cesare, S. (2015). Quantifying stocktwits semantic terms' trading behavior in financial markets: an effective application of decision tree algorithms. Expert Systems with Applications, 42(23):9192–9210.
Beasley, J. E. , Meade, N. , and Chang, T.-J. (2003). An evolutionary heuristic for the index tracking problem. European Journal of Operational Research, 148(3):621–643.
Berry, S. , Mitra, G. , and Sadik, Z. (2019). Improved volatility prediction and trading using stocktwits sentiment data. *White Paper, OptiRisk Systems*.
Canakgoz, N. A. and Beasley, J. E. (2009). Mixed-integer programming approaches for index tracking and enhanced indexation. European Journal of Operational Research, 196(1):384–399.

Dentcheva, D. and Ruszczynski, A. (2006). Portfolio optimization with stochastic dominance constraints. Journal of Banking & Finance, 30(2):433–451.
Fábián, C. I. , Mitra, G. , and Roman, D. (2011). Processing second-order stochastic dominance models using cutting-plane representations. Mathematical Programming, 130(1):33–57.
Fabian, C. I. , Mitra, G. , Roman, D. , and Zverovich, V. (2011). An enhanced model for portfolio choice with SSD criteria: a constructive approach. Quantitative Finance, 11(10):1525–1534.
Feuerriegel, S. and Prendinger, H. (2016). News-based trading strategies. Decision Support Systems, 90:65–74.
Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below-target returns. The American Economic Review, 67(2):116–126.
Hochreiter, R. (2016). Computing trading strategies based on financial sentiment data using evolutionary optimization. In *International Conference on Soft Computing-MENDEL* , pages 181–191. Springer.
Kelly, R. (2009). Twitter Study-August 2009. *San Antionio, TX: Pear Analytics*.
Kroll, Y. (1980). Stochastic Dominance: A Review and Some New Evidence. Research in Finance, 2:163–227.
Kumar, R. , Mitra, G. , and Roman, D. (2008). Long-short portfolio optimisation in the presence of discrete asset choice constraints and two risk measures. *Available at SSRN 1099926*.
Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1):77–91.
Mitra, G. and Mitra, L. (2011). The Handbook of News Analytics in Finance. John Wiley & Sons, Chichester, UK.
Mitra, G. and Yu, X. (2016). The Handbook of Sentiment Analysis in Finance. Optirisk Systems in Collaboration with Albury Books.
Ogryczak, W. and Ruszczynski, A. (1999). From stochastic dominance to mean-risk models: semideviations as risk measures. European Journal of Operational Research, 116(1):33–50.
Ogryczak, W. and Ruszczynski, A. (2001). On consistency of stochastic dominance and mean–semideviation models. Mathematical Programming, 89(2):217–232.
Oliveira, N. , Cortez, P. , and Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Portuguese Conference on Artificial Intelligence* , pages 355–365. Springer.
Oliveira, N. , Cortez, P. , and Areal, N. (2014). Automatic creation of stock market lexicons for sentiment analysis using stocktwits data. In *Proceedings of the 18th International Database Engineering & Applications Symposium* , pages 115–123. ACM.
Pring, M. (2002). Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points. McGraw-Hill Education.
Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. Journal of Banking & Finance, 26(7):1443–1471.
Rockafellar, R. T. , Uryasev, S. , et al. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2:21–42.
Roman, D. , Darby-Dowman, K. , and Mitra, G. (2006). Portfolio construction based on stochastic dominance and target return distributions. Mathematical Programming, 108(2-3):541–569.
Roman, D. and Mitra, G. (2009). Portfolio selection models: a review and new directions. Wilmott Journal: The International Journal of Innovative Quantitative Finance Research, 1(2):69–85.
Roman, D. , Mitra, G. , and Zverovich, V. (2013). Enhanced indexation based on second-order stochastic dominance. European Journal of Operational Research, 228(1):273–281.
Sadik, Z. , Mitra, G. , and Tan, Z. (2019). Asset allocation strategies: enhanced by news. *Whitepaper, OptiRisk Systems.*
Sadik, Z. A. , P. M., Date , and Mitra, G. (2018). News augmented GARCH(1,1) model for volatility prediction. IMA Journal of Management Mathematics, 30(2):165–185.
Sadik, Z. A. , Date, P. M. , and Mitra, G. (2020). Forecasting crude oil futures prices using global macroeconomic news sentiment. IMA Journal of Management Mathematics, 31(2):191–215.
Shi, Y. , Mitra, G. , Arbex-Valle, C. , and Yu, X. (2017). Using social media and news sentiment data to construct a momentum strategy. *Available at SSRN:* https://ssrn.com/abstract=3406075 or http://dx.doi.org/10.2139/ssrn.3406075.
Sprenger, T. O. , Tumasjan, A. , Sandner, P. G. , and Welpe, I. M. (2014). Tweets and trades: the information content of stock microblogs. European Financial Management, 20(5):926–957.
Von Neumann, J. , Morgenstern, O. , and Kuhn, H. W. (2007). Theory of Games and Economic Behavior (commemorative edition). Princeton University Press.
Whitmore, G. A. and Findlay, M. C. (1978). Stochastic Dominance: An Approach to Decision-Making Under Risk. Lexington Books.
Yu, X. (2014). Analysis of New Sentiment and Its Application to Finance. PhD thesis, Brunel University London, Uxbridge, UK.
Yu, X. , Mitra, G. , Arbex-Valle, C. , and Sayer, T. (2015). An Impact Measure for News: Its Use in Daily Trading Strategies. *Available at SSRN:* https://ssrn.com/abstract=2702032 *or* http://dx.doi.org/10.2139/ssrn.2702032.

Asset Allocation Strategies: Enhanced by News

Beasley, J. E. , Meade, N. , and Chang, T.-J. (2003). An evolutionary heuristic for the index tracking problem. European Journal of Operational Research, 148(3):621–643.
Canakgoz, N. A. and Beasley, J. E. (2009). Mixed-integer programming approaches for index tracking and enhanced indexation. European Journal of Operational Research, 196(1):384–399.
De Long, J. B. , Shleifer, A. , Summers, L. H. , and Waldmann, R. J. (1990). Noise trader risk in financial markets. Journal of Political Economy, 98(4):703–738.
Dentcheva, D. and Ruszczynski, A. (2006). Portfolio optimization with stochastic dominance constraints. Journal of Banking & Finance, 30(2):433–451.
DiBartolomeo, D. (2000). The enhanced index fund as an alternative to indexed equity management. *Northfield Information Services, Boston*.
Fabian, C. I. , Mitra, G. , Roman, D. , and Zverovich, V. (2011). An enhanced model for portfolio choice with SSD criteria: a constructive approach. Quantitative Finance, 11(10):1525–1534.

Feuerriegel, S. and Prendinger, H. (2016). News-based trading strategies. Decision Support Systems, 90:65–74.
Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below-target returns. The American Economic Review, 67(2):116–126.
Kantos, C. , Joldzic, D. , Mitra, G. , and Hoang Thi, K. (2022). Comparative Analysis of NLP Approaches for Earnings Calls. *White Paper Available on* https://optirisk-systems.com/publications/whitepapers/.
Kroll, Y. (1980). Stochastic dominance: a review and some new evidence. Research in Finance, 2:163–227.
Kumar, R. , Mitra, G. , and Roman, D. (2008). Long-short portfolio optimisation in the presence of discrete asset choice constraints and two risk measures. *Available at SSRN 1099926*.
Lo, A. W. (1999). The three p's of total risk management. Financial Analysts Journal, 55(1):13–26.
Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1):77–91.
Mitra, G. and Mitra, L. (2011). The Handbook of News Analytics in Finance. John Wiley & Sons, Chichester, UK.
Mitra, G. and Yu, X. (2016). The Handbook of Sentiment Analysis in Finance. Optirisk Systems in Collaboration with Albury Books.
Ogryczak, W. and Ruszczynski, A. (1999). From stochastic dominance to mean-risk models: semideviations as risk measures. European Journal of Operational Research, 116(1):33–50.
Ogryczak, W. and Ruszczynski, A. (2001). On consistency of stochastic dominance and mean–semideviation models. Mathematical Programming, 89(2):217–232.
Pring, M. (2002). Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points. McGraw-Hill Education.
Rockafellar, R. T. and Uryasev, S. (2002). Conditional Value-at-Risk for General Loss Distributions. Journal of Banking & Finance, 26(7):1443–1471.
Rockafellar, R. T. , Uryasev, S. , et al. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2:21–42.
Roman, D. , Darby-Dowman, K. , and Mitra, G. (2006). Portfolio construction based on stochastic dominance and target return distributions. Mathematical Programming, 108(2-3):541–569.
Roman, D. , Mitra, G. , and Zverovich, V. (2013). Enhanced indexation based on second-order stochastic dominance. European Journal of Operational Research, 228(1):273–281.
Sadik, Z. A. P. M., Date , and Mitra, G. (2018). News augmented GARCH(1,1) model for volatility prediction. IMA Journal of Management Mathematics, 30(2):165–185.
Sadik, Z. A. , Date, P. M. , and Mitra, G. (2020). Forecasting crude oil futures prices using global macroeconomic news sentiment. IMA Journal of Management Mathematics, 31(2):191–215.
Scowcroft, A. and Satchell, S. (2003). Advances in Portfolio Construction and Implementation. Elsevier.
Vickson, R. (1975). Stochastic dominance tests for decreasing absolute risk aversion. i. discrete random variables. Management Science, 21(12):1438–1446.
Von Neumann, J. , Morgenstern, O. , and Kuhn, H. W. (2007). Theory of Games and Economic Behavior (commemorative edition). Princeton University Press.
Whitmore, G. A. and Findlay, M. C. (1978). Stochastic Dominance: An Approach to Decision-Making Under Risk. Lexington Books.
Yu, X. (2014). Analysis of New Sentiment and Its Application to Finance. PhD thesis, Brunel University London, Uxbridge, UK.
Yu, X. , Mitra, G. , Arbex-Valle, C. , and Sayer, T. (2015). An Impact Measure for News: Its Use in Daily Trading Strategies. *Available at SSRN:* https://ssrn.com/abstract=2702032 *or* http://dx.doi.org/10.2139/ssrn.2702032.

Extracting Structured Datasets from Textual Sources – Some Examples

Ahelegbey, D. F. , Cerchiello, P. , and Scaramozzino, R. (2022). Network based evidence of the financial impact of Covid-19 pandemic. International Review of Financial Analysis, 81:102101.
Alpha, E. (2021). Eagle alpha's 1st annual alternative data report – 2021. https://www.eaglealpha.com/2021/11/18/eagle-alphas-1st-annual-alternative-data-report-2021/.
Alternative Data Org (2022). Alternative data database. https://alternativedata.org/. Accessed: 2022-06-23.
Bouchaud, J.-P. (2022). The inelastic market hypothesis: a microstructural interpretation. Quantitative Finance, pages 1–11.
Brain (2022). Brain website. https://braincompany.co. Accessed: 2022-06-23.
Cohen, L. , Malloy, C. , and Nguyen, Q. (2020). Lazy prices. The Journal of Finance, 75(3):1371–1415.
Confessore, N. (2018). Cambridge analytica and facebook: The scandal and the fallout so far. The New York Times, 4:2018.
Frankel, R. M. , Jennings, J. N. , and Lee, J. A. (2017). Using natural language processing to assess text usefulness to readers: the case of conference calls and earnings prediction. *Available at SSRN 3095754*.
Gabaix, X. and Koijen, R. S. (2021). In search of the origins of financial fluctuations: the inelastic markets hypothesis. Technical report, National Bureau of Economic Research.
Gunning, R. et al. (1952). Technique of Clear Writing. *McGraw-Hill* , pages 36–37.
Hanicova, D. , Kalus, F. , and Vojtko, R. (2021). How to use lexical density of company filings. *Available at SSRN 3921091*.
Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. Working papers/Lund University, Department of Linguistics and Phonetics, 53:61–79.
Kimbrough, M. D. (2005). The effect of conference calls on analyst and market underreaction to earnings announcements. The Accounting Review, 80(1):189–219.
Kincaid, J. P. , Fishburne Jr, R. P. , Rogers, R. L. , and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1):35–65.
Malkiel, B. G. (2003). The efficient market hypothesis and its critics. Journal of Economic Perspectives, 17(1):59–82.
Matsumoto, D. , Pronk, M. , and Roelofsen, E. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. The Accounting Review, 86(4):1383–1414.

Mayew, W. J. and Venkatachalam, M. (2012). The power of voice: managerial affective states and future firm performance. The Journal of Finance, 67(1):1–43.
Mitra, G. and Mitra, L. (2011). The Handbook of News Analytics in Finance. John Wiley & Sons.
Mitra, G. and Yu, X. (2016). Handbook of Sentiment Analysis in Finance. Optirisk Systems in Collaboration with Albury Books.
Padysak, M. (2020). The positive similarity of company filings and the cross-section of stock returns. *Available at SSRN 3690461*.
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3):1139–1168.
Wan, X. , Yang, J. , Marinov, S. , Calliess, J.-P. , Zohren, S. , and Dong, X. (2021). Sentiment correlation in financial news networks and associated market movements. Scientific Reports, 11(1):1–12.
Wikipedia (2019). Natural language processing. https://en.wikipedia.org/wiki/Natural_language_processing. Accessed 2022-06-23.

Comparative Analysis of NLP Approaches for Earnings Calls

(2019). Factset research systems 2019. https://investor.factset.com/static-files/83ca372c-8283-419f-a8c7-40c8e0daab51.
Araci, D. F. and Genc, Z. (2019). Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
Chen, J. (2021). Earnings call. https://www.investopedia.com/terms/e/earnings-call.asp.
Govindaraj, S. , Livnat, J. , Savor, P. G. , and Zhao, C. (2013). Large price changes and subsequent returns. *Journal Of Investment Management*.
Jha, V. , Blaine, J. , and Montague, W. (2015). Finding value in earnings transcripts data with alphasense. Extractalpha.com.
Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1):35–65.
Short, J. (2016). Societe generale cross asset research/equity quant. https://silo.tips/download/societe-generale-cross-asset-research.
Tortoriello, R. (2018). Their Sentiments Exactly: Sentiment Signal Diversity Creates Alpha Opportunity. *S&P Global*.

Sensors Data

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives, 15(4):69–85.
Curran, K. and Orr, J. (2011). Integrating geolocation into electronic finance applications for additional security. International Journal of Electronic Finance, 5(3):272–285.
Daglis, T. , Konstantakis, K. N. , Michaelides, P. G. , and Papadakis, T. E. (2020). The forecasting ability of solar and space weather data on NASDAQ's finance sector price index volatility. Research in International Business and Finance, 52:101–147.
Fisher, R. A. (1925). III. The influence of rainfall on the yield of wheat at Rothamsted. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, 213(402-410):89–142.
FlyXO (2022). Tracking private jets for investment tips! https://flyxo.com/blog/tracking-private-jets-investment-tips/.
Han, S. , Ahn, D. , Park, S. , Yang, J. , Lee, S. , Kim, J. , Yang, H. , Park, S. , and Cha, M. (2020). Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* , pages 2970–2979.
Henderson, J. V. , Storeygard, A. , and Weil, D. N. (2012). Measuring economic growth from outer space. American Economic Review, 102(2):994–1028.
Hernandez-Sosa, D. , Castrillón-Santana, M. , and Lorenzo-Navarro, J. (2011). Multi-sensor people counting. In *Iberian Conference on Pattern Recognition and Image Analysis* , pages 321–328. Springer.
Jain, H. and Jain, R. (2017). Big data in weather forecasting: Applications and challenges. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)* , pages 138–142. IEEE.
Merkel, B. , Phillips, R. A. , Descamps, S. , Yoccoz, N. G. , Moe, B. , and Strom, H. (2016). A probabilistic algorithm to process geolocation data. Movement Ecology, 4(1):1–11.
Mundhenk, T. N. , Konjevod, G. , Sakla, W. A. , and Boakye, K. (2016). A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision* , pages 785–800. Springer.
National Centers for Environmental Information - National Oceanic and Atmospheric Administration . https://www.ncdc.noaa.gov/cdo-web/ Accessed: 2023-05-27.
Schwalbert, R. A. , Amado, T. , Corassa, G. , Pott, L. P. , Prasad, P. V. , and Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern brazil. Agricultural and Forest Meteorology, 284:107886.
Virtue, G. (1929). The tariff on animal and vegetable oils. JSTOR.

Media Sentiment Momentum

Barber, B. M. , Huang, X. , Odean, T. , and Schwarz, C. (2021). Attention induced trading and returns: evidence from RobinHood users. *Available at SSRN:3715077*.
Barberis, N. , Shleifer, A. , and Vishny, R. (1998). A model of investor sentiment. Journal of Financial Economics, 49(3):307–343.
Beckers, S. (2018). Do social media trump news? The relative importance of social media and news based sentiment for market timing. The Journal of Portfolio Management, 45(2):58–67.

Bernard, V. L. and Thomas, J. K. (1989). Post-earnings-announcement drift: delayed price response or risk premium? Journal of Accounting Research, 27:1–36.
Cahan, R. B. Y. Y. S. , (2018a). Big Data: Harnessing News and Social Media to Improve Our Timing. *Empirical Research Partners*.
Cahan, R. B. Y. Y. S. , (2018b). Big Data: Media Sentiment: Useful Around the World. *Empirical Research Partners*.
Chen, J. Z. , Lobo, G. J. , and Zhang, J. H. (2017). Accounting quality, liquidity risk, and post-earnings-announcement drift. Contemporary Accounting Research, 34(3):1649–1680.
Connell, D. (2015). Institutional investing: how social media informs and shapes the investing process. https://www.greenwich.com/asset-management/institutional-investing-how-social-mediainforms-and-shapes-investing-process.
Daniel, K. D. (2011). Momentum crashes. *Columbia Business School Research Paper. Available at SSRN: 1914673*, pages 11–13.
Daniel, K. D. , Hirshleifer, D. A. , and Subrahmanyam, A. (1998). A theory of overconfidence, self-attribution, and security market under- and over-reactions. Journal of Finance, 53(6):1839–1885.
de la Vega, J. (1688). Confusion de confusiones. Colchis Books, 13:289–293.
Dougal, C. , Engelberg, J. , Garcia, D. , and Parsons, C. A. (2012). Journalists and the stock market. The Review of Financial Studies, 25(3):639–679.
Ehsani, S. and Linnainmaa, J. T. (2022). Factor momentum and the momentum factor. The Journal of Finance, 77(3):1877–1919.
Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. The Journal of Finance, 51(1):55–84.
Fischer, D. (2012). Investor underreaction to earnings surprises and overreaction to product news in the drug industry. The Journal of Business and Economic Studies, 18(2):82.
Giglio, S. and Shue, K. (2014). No news is news: do markets underreact to nothing? The Review of Financial Studies, 27(12):3389–3440.
Glaser, F. and Risius, M. (2018). Effects of transparency: analyzing social biases on trader performance in social trading. Journal of Information Technology, 33(1):19–30.
Griffin, J. M. , Ji, X. , and Martin, J. S. (2005). Global momentum strategies. The Journal of Portfolio Management, 31(2):23–39.
Grinblatt, M. , Titman, S. , and Wermers, R. (1995). Momentum investment strategies, portfolio performance, and herding: a study of mutual fund behavior. The American Economic Review, 85(5):1088–1105.
Hirshleifer, D. (2020). Presidential address: social transmission bias in economics and finance. The Journal of Finance, 75(4):1779–1831.
Hirshleifer, D. A. , Myers, J. N. , Myers, L. A. , and Teoh, S. H. (2008). Do individual investors cause post-earnings announcement drift? Direct evidence from personal trades. The Accounting Review, 83(6):1521–1550.
Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. The Journal of Finance, 54(6):2143–2184.
Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. The Journal of Finance, 48(1):65–91.
Lee, J. (2020). Momentum trade plunges the most on record in rotation frenzy. https://www.bloomberg.com/news/articles/2020-11-09/momentum-trade-plunges-the-most-on-record-in-rotation-frenzy.
Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*.
Lim, S. , Lane, J. , and Uzzi, B. (2020). Biased information transmission in investor social networks: evidence from professional traders. In Academy of Management Proceedings, volume 2020. Academy of Management Briarcliff Manor, NY 10510.
Liu, C. , L. A.. P. R. (2020). StarMine MarketPsych Media Sentiment Model Product Whitepaper. *Refinitiv*.
Lo, A. W. (2017). Adaptive markets: financial evolution at the speed of thought. In Adaptive Markets, pages 281–282. Princeton University Press.
Mendenhall, R. R. (2004). Arbitrage risk and post-earnings-announcement drift. The Journal of Business, 77(4):875–894.
Peterson, R. & Fischkin, E. (2013). Thomson Reuters MarketPsych Analytics (TRMA) White Paper. *Thomson Reuters*.
Pfarrer, M. D. , Pollock, T. G. , and Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. Academy of Management Journal, 53(5):1131–1152.
Rantala, V. (2019). How do investment ideas spread through social interaction? evidence from a ponzi scheme. The Journal of Finance, 74(5):2349–2389.
Rozin, P. and Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review, 5(4):296–320.
Tesser, A. and Rosen, S. (1975). The reluctance to transmit bad news. In Advances in Experimental Social Psychology, volume 8, pages 193–232. Elsevier.
Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? The Review of Financial Studies, 24(5):1481–1512.
Tetlock, P. C. , Saar-Tsechansky, M. , and Macskassy, S. (2008). More than words: quantifying language to measure firms' fundamentals. The Journal of Finance, 63(3):1437–1467.
Welch, I. (2022). The wisdom of the robinhood crowd. The Journal of Finance, 77(3):1489–1527.

Defining Market States with Media Sentiment

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected Papers of Hirotugu Akaike, pages 199–213. Springer.
Andersen, T. G. , Bollerslev, T. , Diebold, F. X. , and Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets. Journal of International Economics, 73(2):251–277.
Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. Annual Review of Financial Economics, 4(1):313–337.
Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. The Journal of Finance, 61(4):1645–1680.
Bali, T. G. , Demirtas, K. O. , Levy, H. , and Wolf, A. (2009). Bonds versus stocks: investors' age and risk taking. Journal of Monetary Economics, 56(6):817–830.
Black, F. (1976). Studies of stock market volatility changes. *1976 Proceedings of the American Statistical Association Business and Economic Statistics section* . American Statistical Association.

Chaves, D. , Hsu, J. , Li, F. , and Shakernia, O. (2011). Risk parity portfolio vs. other asset allocation heuristic portfolios. The Journal of Investing, 20(1):108–118.
Christie, A. A. (1982). The stochastic behavior of common stock variances: value, leverage and interest rate effects. Journal of Financial Economics, 10(4):407–432.
Ciner, C. , Gurdgiev, C. , and Lucey, B. M. (2013). Hedges and safe havens: an examination of stocks, bonds, gold, oil and exchange rates. International Review of Financial Analysis, 29:202–211.
Jiang, F. , Lee, J. , Martin, X. , and Zhou, G. (2019). Manager sentiment and stock returns. Journal of Financial Economics, 132(1):126–149.
Neath, A. A. and Cavanaugh, J. E. (2012). The bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2):199–203.
Pedregosa, F. , Varoquaux, G. , Gramfort, A. , Michel, V. , Thirion, B. , Grisel, O. , Blondel, M. , Prettenhofer, P. , Weiss, R. , Dubourg, V. , Vanderplas, J. , Passos, A. , Cournapeau, D. , Brucher, M. , Perrot, M. , and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.
Schaller, H. and Norden, S. V. (1997). Regime switching in stock market returns. Applied Financial Economics, 7(2):177–191.
Shiller, R. J. (2003). From efficient markets theory to behavioral finance. Journal of Economic Perspectives, 17(1):83–104.

A Quantitative Metric for Corporate Sustainability

Amiraslani, H. , Lins, K. , Servaes, H. , and Tamayo, A. The Bond Market Benefits of Corporate Social Capital. *European Corporate Governance Institute Working Paper*.
Belev, E. and diBartolomeo, D. (2019). Finance meets macroeconomics: a structural model of sovereign credit risk. In *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 4: Contingent Claims Approach for Banks and Sovereign Debt* , pages 433–461. World Scientific.
Black, F. and Cox, J. C. (1976). Valuing corporate securities: some effects of bond indenture provisions. The Journal of Finance, 31(2):351–367.
Bragdon, J. H. and Marlin, J. (1972). Is pollution profitable. Risk Management, 19(4):9–18.
DiBartolomeo, D. (2010). Equity risk, credit risk, default correlation, and corporate sustainability. The Journal of Investing, 19(4):128–133.
Dyer, S. and Zieff, W. (2017). Performance of ESG Sustainable Investing and Minimum Variance Strategies. https://www.northinfo.com/documents/771.pdf.
Friedman, M. (1970). A Friedman doctrine - The Social Responsibility of Business Is to Increase Its Profits. The New York Times. https://www.nytimes.com Accessed: 2023-05-27.
Henke, H.-M. (2016). The effect of social screening on bond mutual fund performance. Journal of Banking & Finance, 67:69–84.
Hotelling, H. (1931). The economics of exhaustible resources. Journal of Political Economy, 39(2):137–175.
Jobson, J. D. and Korkie, B. M. (1981). Performance hypothesis testing with the sharpe and treynor measures. Journal of Finance, 36(4):889–908.
Kantos, C. and DiBartolomeo, D. (2020). How the pandemic taught us to turn smart beta into real alpha. Journal of Asset Management, 21(7):581–590.
Kanuri, S. and McLeod, R. W. (2016). Sustainable competitive advantage and stock performance: the case for wide moat stocks. Applied Economics, 48(52):5117–5127.
Khan, M. , Serafeim, G. , and Yoon, A. (2016). Corporate sustainability: first evidence on materiality. The Accounting Review, 91(6):1697–1724.
Kurtz, L. (2012). Macro factors in corporate governance. Northfield Asia Research Seminar.
Kurtz, L. and DiBartolomeo, D. (1996). Socially screened portfolios: an attribution analysis of relative performance. The Journal of Investing, 5(3):35–41.
Kurtz, L. et al. (2011). The long-term performance of a social investment universe. The Journal of Investing, 20(3):95–102.
Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. Journal of Empirical Finance, 15(5):850–859.
Leland, H. E. (1994). Corporate debt value, bond covenants, and optimal capital structure. The Journal of Finance, 49(4):1213–1252.
Leland, H. E. and Toft, K. B. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. The Journal of Finance, 51(3):987–1019.
Memmel, C. (2003). Performance hypothesis testing with the sharpe ratio. *Available at SSRN 412588*.
Merton, R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates. The Journal of finance, 29(2):449–470.
Messmore, T. (1995). Variance drain. Journal of Portfolio Management, 21(4):104.
Plantinga, A. and de Groot, S. (2001). Risk Adjusted Performance Measures and Implied Risk Attitudes. *Capital Markets eJournal*.
Ricketts, D. (2021). Ex-BlackRock sustainable CIO on how ESG risks turning into a mis-selling scandal. Financial News. https://www.fnlondon.com/articles/ex-blackrock-sustainable-cio-on-how-esg-risks-turning-into-a-mis-selling-scandal-20210914 Accessed: 2023-05-27.
Yaksick, R. (1995). Expected optimal exercise time of a perpetual american option: a closed-form solution. In Advances in Stochastic Modelling and Data Analysis, pages 29–56. Springer.

Hot Off the Press: Predicting Intraday Risk and Liquidity with News Analytics

Allen, D. E. , McAleer, M. J. , and Singh, A. K. (2015). Machine News and Volatility: The Dow Jones Industrial Average and the TRNA Real-Time High-Frequency Sentiment Series. *Academic Press San Diego* , pages 327–344.
Barunik, J. , Chen, C. Y.-H. , and Vecer, J. (2019). Sentiment-driven stochastic volatility model: a high-frequency textual tool for economists. *arXiv preprint arXiv:1906.00059*.

Boudoukh, J. , Feldman, R. , Kogan, S. , and Richardson, M. (2019). Information, trading, and volatility: evidence from firm-specific news. The Review of Financial Studies, 32(3):992–1033.
Caporin, M. and Poli, F. (2017). Building news measures from textual data and an application to volatility forecasting. Econometrics, 5(3):35.
Clements, A. , Fuller, J. , Papalexiou, V. , et al. (2015). Public news flow in intraday component models for trading activity and volatility. Technical report, National Centre for Econometric Research.
Conrad, C. and Engle, R. F. (2021). Modelling volatility cycles: the MF2-GARCH model. *Working Paper series 21-05, Rimini Centre for Economic Analysis*.
Creamer, G. G. (2015). Can a corporate network and news sentiment improve portfolio optimization using the Black–Litterman model? Quantitative Finance, 15(8):1405–1416.
Cui, B. and Gozluklu, A. E. (2021). News and trading after hours. *Available at SSRN 3796812*.
Engle, R. F. , Giglio, S. , Kelly, B. , Lee, H. , and Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3):1184–1216.
Gross-Klussmann, A. and Hautsch, N. (2011). When machines read the news: using automated text analytics to quantify high frequency news-implied market reactions. Journal of Empirical Finance, 18(2):321–340.
Ho, K.-Y. , Shi, Y. , and Zhang, Z. (2013). How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches. The North American Journal of Economics and Finance, 26:436–456.
Kalev, P. S. , Liu, W.-M. , Pham, P. K. , and Jarnecic, E. (2004). Public information arrival and volatility of intraday stock returns. Journal of Banking & Finance, 28(6):1441–1467.
Mitra, G. , Erlwein-Sayer, C. , Valle, C. A. , and Yu, X. (2018). Using market sentiment to enhance second-order stochastic dominance trading models. In High-Performance Computing in Finance, pages 25–48. Chapman and Hall/CRC.
Mitra, L. , Mitra, G. , and Dibartolomeo, D. (2009). Equity portfolio risk estimation using market information and sentiment. Quantitative Finance, 9(8):887–895.
Rahimikia, E. and Poon, S.-H. (2020). Big data approach to realised volatility forecasting using har model augmented with limit order book and news. *Available at SSRN 3684040*.
Riordan, R. , Storkenmaier, A. , Wagener, M. , and Zhang, S. S. (2013). Public information arrival: price discovery and liquidity in electronic limit order markets. Journal of Banking & Finance, 37(4):1148–1159.
Shi, Y. , Ho, K.-Y. , and Liu, W.-M. (2016). Public information arrival and stock return volatility: evidence from news sentiment and markov regime-switching approach. International Review of Economics & Finance, 42:291–312.
Song, S. and Bhattacharyya, R. (2021). The intraday impact of macroeconomic news on price moves of financial instruments. *Available at SSRN 3798844*.
Tetlock, P. C. (2010). Does public financial news resolve asymmetric information? The Review of Financial Studies, 23(9):3520–3557.
Uctum, R. , Renou-Maissant, P. , Prat, G. , and Lecarpentier-Moyal, S. (2017). Persistence of announcement effects on the intraday volatility of stock returns: evidence from individual data. Review of Financial Economics, 35:43–56.

Exogenous Risks, Alternative Data Implications for Strategic Asset Allocation – Multi-Subordination Levy Processes Approach

Basel Committee et al. (2014). Fundamental review of the trading book: A revised market risk framework. https://www.bis.org/publ/bcbs265.pdf.
Bilarev, T. , Mitov, G. , Racheva-Iotova, B. , and Stefanov, I. (n.d.). Multi-asset class portfolio construction. Technical report, FACTSET.
Bingler, J. A. and Colesanti Senni, C. (2020). Taming the green swan: how to improve climate-related financial risk assessments. *Available at SSRN 3795360*.
FactSet . (n.d.). ESG Data and Analytics from Truvalue Labs. Technical report, FactSet.
GEOQUANT (2020). The year ahead 2020 data-driven predictions for global investors. Technical report, GEOQUANT.
Giese, G. , Nagy, Z. , and Rauis, B. (2021). Foundations of climate investing: how equity markets have priced climate-transition risks. The Journal of Portfolio Management, 47(9):35–53.
Karagozoglu, A. K. (2021). Novel risks: a research and policy overview. The Journal of Portfolio Management, 47(9):11–34.
Kazdin, J. , Schwaiger, K. , Wendt, V.-S. , and Ang, A. (2021). Climate alpha with predictors also improving company efficiency. The Journal of Impact and ESG Investing, 2(2):35–56.
Ken-Iti, S. (1999). Levy Processes and Infinitely Divisible Distributions. Cambridge University Press.
Rachev, S. , Fabozzi, F. J. , Racheva-Iotova, B. , and Shirvani, A. (2017). Option pricing with greed and fear factor: the rational finance approach. *arXiv preprint arXiv:1709.08134*.
Rachev, S. T. , Kim, Y. S. , Bianchi, M. L. , and Fabozzi, F. J. (2011). Financial Models with Lévy Processes and Volatility Clustering. John Wiley & Sons.
Rachev, S. T. and Mittnik, S. (2000). Stable Paretian models in finance, volume 7. Wiley, ISBN: 978-0-471-95314-2.
Rachev, S. T. , Racheva-Iotova, B. , Stoyanov, S. V. , and Fabozzi, F. J. (2010). Risk management and portfolio optimization for volatile markets. In Handbook of Portfolio Construction, pages 493–508. Springer.
Racheva-Iotova, B. (2020a). Back-testing studies on the factset short-term risk model during the covid-19 outbreak. *Factset Insight Article.*
Racheva-Iotova, B. (2020b). Wealth Management Next Frontiers-The Inevitable Need to Meet Behavioral and Quantitative Approaches. In Handbook of Applied Investment Research, pages 479–509. World Scientific.
Racheva-Iotova, B. , Mitov, I. , Mossessian, D. , and Bodurov, V. (2018). Factset fat-tail multi-asset class model - introducing turbulence adjusted risk. Technical report, FACTSET.
Radev, N. , Mitov, G. , and Racheva-Iotova, B. (n.d.). Minimum tail risk portfolio construction. Technical report, FACTSET.
Roncalli, T. , Le Guenedal, T. , Lepetit, F. , Roncalli, T. , and Sekine, T. (2021). The market measure of carbon risk and its impact on the minimum variance portfolio. The Journal of Portfolio Management, 47(9):54–68.

Samorodnitsky, G. and Taqqu, M. S. (2017). Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance: Stochastic Modeling. Routledge.
Schoutens, W. (2003). Lévy Processes in Finance: Pricing Financial Derivatives. Wiley Online Library.
Schoutens, W. and Cariboni, J. (2010). Lévy Processes in Credit Risk. John Wiley & Sons.
Stein, M. , Rachev, S. T. , and Stoyanov, S. (2009). R ratio optimization with heterogeneous assets using genetic algorithm. Investment Management and Financial Innovations, 6(3-1):117–125.
Tokat-Acikel, Y. , Aiolfi, M. , Johnson, L. , Hall, J. , and Jin, J. Y. (2021). Top-down portfolio implications of climate change. The Journal of Portfolio Management, 47(9):69–91.
Watson, W. T. (2019). Total portfolio approach (tpa) - a global asset owner study into current and future asset allocation practices. Technical report, ThinkingAhead Institute.

ESG Controversies and Stock Returns

Belsom, T. and Lake, L. (2021). ESG factors and equity returns – a review of recent industry research. *PRI Blog*.
Bloomberg Intelligence (2021). ESG assets may hit $53 trillion by 2025, a third of global AUM. *Bloomberg*.
Bouchaud, J.-P. , Matacz, A. , and Potters, M. (2001). Leverage effect in financial markets: the retarded volatility model. Physical Review Letters, 87(22):228701.
Brandon, R. G. , Krueger, P. , and P. S. Schmidt (2021). ESG rating disagreement and stock returns. Financial Analysts Journal, 77(4):104–127.
Chiappini, H. , Vento, G. , and De Palma, L. (2021). The impact of COVID-19 lockdowns on sustainable indexes. Sustainability, 13(4).
Clark, G. , Feiner, A. , and Viehs, M. (2015). From the stockholder to the stakeholder: how sustainability can drive financial outperformance. *Arabesque*.
Compact, U. (2004). Who cares wins: connecting financial markets to a changing world. *International Finance Corporation*.
Dorfleitner, G. , Halbritter, G. , and Nguyen, M. (2015). Measuring the level and risk of corporate responsibility–an empirical comparison of different esg rating approaches. Journal of Asset Management, 16(7):450–466.
Friede, G. , Busch, T. , and Bassen, A. (2015). ESG and financial performance: aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4):210–233.
Friedman, M. (1962). Capitalism and freedom. University of Chicago Press.
Friedman, M. (1970). A Friedman doctrine - the social responsibility of business is to increase its profits. *The New York Times*.
Hale, J. (2018). Sustainable funds, U.S. landscape report. *Morningstar Research*.
Lester, A. and McKnett, C. (2018). Harnessing ESG as an alpha source in active quantitative equities. *State Street Global Advisors*.
Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*.
Li, F. and Polychronopoulos, A. (2020). What a difference an ESG ratings provider makes. *Research Affiliates*.
Mackintosh, J. (2018). Is Tesla or Exxon more sustainable? It depends whom you ask. *Wall Street Journal*.
McCrum, D. (2019). Executive at Wirecard suspected of using forged contracts. *Financial Times*.
Refinitiv (2021). Refinitiv Marketpsych ESG Analytics: Quantifying sustainability in global news and social media.
Schwab, K. (2021). Stakeholder Capitalism: A Global Economy that Works for Progress, People and Planet. *John Wiley & Sons* .
Subramanian, S. , Kabra, M. , Chopra, S. , Yeo, J. , Strzelinska, P. , and Huang, L. (2019). ESG matters ESG from A to Z: a global primer. *Bank of America Merrill Lynch Research*.
Uhl, M. W. (2014). Reuters sentiment and stock returns. Journal of Behavioral Finance, 15(4):287–298.
Utkarsh, A. and van Broekhuizen, R. (2021). Responsible investing and asset allocation. *NN Investment Partners*.
Wong, G. (2021). Quantitative investing gets sentimental. *AGF Investments*.

Oil and Gas Drilling Waste – A Material Externality

American Coal Ash Association . (n.d.). Fly ash use in concrete increases slightly as overall coal ash recycling rate declines. https://acaa-usa.org/wp-content/uploads/coal-combustion-products-use/Coal-Ash-Production-and-Use.pdf.
American Petroleum Institute (2000). Overview of exploration and production waste volumes and waste management practices in the united states based on api survey of onshore and coastal exploration and production operations for 1995 and api survey of natural gas processing plants for 1995. ICF Consulting.
Bagnoli, M. , Godwin, T. , and Watts, S. G. (2021). On the real effects of changes in definitions of materiality. *Available at SSRN 3973688*.
Bartels, W. (2014). Sustainable insight: the essentials of materiality assessment. https://assets.kpmg/content/dam/kpmg/pdf/2014/10/materiality-assessment.pdf.
Billingsley, P. and Harrington, J. (n.d.). Preventing new groundwater pollution from old oilfield areas' oklahoma corporation commission and central oklahoma governments. https://cese.utulsa.edu/wp-content/uploads/2017/06/IPEC-2013-PREVENTING-GROUNDWATER-POLLUTION-FROM-OLD-OILFIELD-AREAS.pdf. pdf. Accessed: 2022-01-02.
Global Reporting Initiative . (n.d.-a). GRI 11: Oil and Gas Sector 2021, Global Sustainability Standards Board.
Global Reporting Initiative . (n.d.-b). Our mission and history. https://www.globalreporting.org/about-gri/mission-history.
Hale, A. (2017). Overview of OBM cuttings remediation methods and technologies. https://www.aade.org/application/files/2815/7365/8670/Arthur_Hale_Aramco-_Overview_of_OBM_Cuttings_Remediation_Methods_and_Technologies_11-162017.pdf.
Helbling, T. (2020). Externalities: prices do not capture all costs. *International Monetary Fund, Finance & Development*.
Hovorka, S. (1997). Salt Cavern Studies – Regional Map of Salt Thickness in the Midland Basin, Contract Number DE-AF22-96BC14978. Technical report, U.S. Department of Energy, Bureau of Economic Geology, The University of Texas.

Institute, C. F . (n.d.). Materiality thresholds in audits. https://corporatefinanceinstitute.com/resources/knowledge/accounting/materiality-threshold-in-audits. Accessed: 2022-01-22.
IPEICA (2009). Drilling fluids and health risk management.
IPIECA . (n.d.-a). About us. https://www.ipieca.org/about-us. Accessed: 2022-01-29.
IPIECA . (n.d.-b). Performance reporting. https://www.ipieca.org/our-work/sustainability/performance-reporting. Accessed: 2022-01-29.
IPIECA . (n.d.-c). Sustainability reporting guidance for the oil and gas industry module 4 environment. https://www.ipieca.org/media/5115/ipieca_sustainability-guide-2020_mod4-env.pdf. Accessed: 2022-01-19.
IPIECA . (n.d.-d). Worldwide drilling density. https://www.researchgate.net/figure/World-oil-and-gas-well-distribution-and-density-courtesy-of-IHS-Energy_fig16_329707783/. Accessed: 2022-01-22.
Jha, P. K. , Mahto, V. , and Saxena, V. (2014). Emulsion based drilling fluids: an overview. International Journal of ChemTech Research, 6(4):2306–2315.
Johnston, J. E. , Lim, E. , and Roh, H. (2019). Impact of upstream oil extraction and environmental public health: a review of the evidence. Science of the Total Environment, 657:187–199.
Keller, R. P. (2017). Evaluating Drivers of Liability, Risk, and Cost While Enhancing Sustainability for Drilling Waste. Paper presented at the American Association of Drilling Engineers National Technical Conference and Exhibition, Houston, Texas.
Lodungi, J. F. , Alfred, D. , Khirulthzam, A. , Adnan, F. , and Tellichandran, S. (2016). A review in oil exploration and production waste discharges according to legislative and waste management practices perspective in malaysia. International Journal of Waste Resources, 7(1):260.
Murphy, E. (2014). Reserve Pit and Brine Pond Studies in North Dakota. *Energy Development and Transmission Committee Senator Rich Wardner, Chair Minot, ND April*, 8.
Nations, U. (n.d.). Goal 12: Ensure sustainable consumption and production patterns, goal 12 targets. https://www.un.org/sustainabledevelopment/sustainable-consumption-production/. Accessed: 2022-01-07.
Norwegian Environment Agency . (n.d.). PFAS in mining and petroleum industry – use, emissions and alternatives.
Pragst, F. , Stieglitz, K. , Runge, H. , Runow, K.-D. , Quig, D. , Osborne, R. , Runge, C. , and Ariki, J. (2017). High concentrations of lead and barium in hair of the rural population caused by water pollution in the thar jath oilfields in south sudan. Forensic Science International, 274:99–106.
Programme, U. N. E. (n.d.). Goal 12: Sustainable consumption and production goal 12: sustainable consumption and production. https://www.unep.org/explore-topics/sustainable-development-goals/why-do-sustainable-development-goals-matter/goal-12/. Accessed: 2022-01-08.
Ramirez Jr, P. (2009). *Reserve Pit Management: Risks to Migratory Birds. Cheyenne (WY): US Fish and Wildlife Service*.
Reader, H. P. (n.d.). The North Dakota Salt Lands. https://hpr1.com/index.php/feature/news/the-north-dakota-salt-lands. Accessed: 2022-01-10.
Reis, J. (1993). An Overview of the Environmental Issues Facing the Upstream Petroleum Industry. In *SPE Annual Technical Conference and Exhibition* . Society of Petroleum Engineers.
SASB . (n.d.-a). Find your industry, extractives & minerals processing. https://www.sasb.org/standards/materiality-finder/find. Accessed: 2022-01-29.
SASB . (n.d.-b). Oil & gas – exploration & production, disclosure topics. https://www.sasb.org/standards/materiality-finder/find/?industry=EM-EP. Accessed: 2022-01-29.
SASB . (n.d.-c). Oil & gas – services, disclosure topics. https://www.sasb.org/standards/materiality-finder/find/?industry=EM-SV. Accessed: 2022-01-29.
SASB . (n.d.-d). The research scope of sasb standards encompasses a range of sustainability issues, environment. https://www.sasb.org/standards/materiality-finder/?lang=en-us. Accessed: 2022-01-29.
SASB (2020). Sasb standards & other esg frameworks. https://www.sasb.org/about/sasb-and-other-esg-frameworks. Accessed 2022-01-29.
Scott, B. (2010). Opportunities and obstacles to reducing the environmental footprint of natural gas development in the Uinta Basin' Colorado Law Scholarly Commons. https://scholar.law.colorado.edu/cgi/viewcontent.cgi?article=1004&context=reducing-environmental-footprint-of-natural-gas-development-in-uintah-basin.
Stuckman, M. , Thomas, C. , Lopano, C. , and Hakala, A. (2015). Leaching characteristics of drill cuttings from unconventional gas reservoirs. In *Unconventional Resources Technology Conference* , *San Antonio, Texas* , *20-22 July 2015*, pages 1301–1308. Society of Exploration Geophysicists, American Association of Petroleum.
Swiedler, E. W. , Muehlenbachs, L. A. , Chu, Z. , Shih, J.-S. , and Krupnick, A. (2019). Should solid waste from shale gas development be regulated as hazardous waste? Energy Policy, 129:1020–1033.
Thuermer Jr., A. (2018). Pavillion water experts fault leaky gas, unlined pits. https://wyofile.com/pavillion-water-experts-fault-leaky-gas-wells-unlined-pits/. Accessed: 2022-01-28.
Tuncan, A. , Tuncan, M. , and Koyuncu, H. (1996). Reuse of stabilized petroleum drilling wastes as sub-bases. *AA Balkema* .
United States Department of Energy , E. I. A. (n.d.). Today in energy drilling and completion improvement support improvement in permian basin hydrocarbons production. https://www.eia.gov/todayinenergy/detail.php?id=50016. Accessed: 2022-01-08.
U.S. Department of the Interior , E. P. A. (2018). Inventory of U.S. greenhouse gas emissions and sinks 1990–2016: Abandoned oil and gas wells. https://www.epa.gov/sites/default/files/2018-01/documents/2018_complete_report.pdf. Accessed: 2022-01-09.
U.S. Department of the Interior , E. P. A. O. o. R. C. and Recovery (2019). Management of exploration and production wastes: factors informing a decision on the need for regulatory action.
U.S. Department of the Interior , Environmental Protection Agency, Office of Solid Waste and Emergency Response (1987). Report to Congress, Management of Wastes from the Exploration, Development, and Production of Crude Oil, Natural Gas, and Geothermal Energy. 1.

Ashwin Kumar, N. , Smith, C. , Badis, L. , Wang, N. , Ambrosy, P. , and Tavares, R. (2016). ESG factors and risk-adjusted performance: A new quantitative model. Journal of Sustainable Finance & Investment, 6(4):292–300.

Barroso, P. and Santa-Clara, P. (2015). Momentum has its moments. Journal of Financial Economics, 116(1):111–120.

Butt, H. A. , Kolari, J. W. , and Sadaqat, M. (2020). Revisiting momentum profits in emerging markets. Pacific-Basin Finance Journal, 65:101–486.

Coqueret, G. , Stiernegrip, S. , Morgenstern, C. , Kelly, J. , Frey-Skott, J. , and Osterberg, B. (2021). Boosting esg-based optimization with asset pricing characteristics. *Available at SSRN 3877242*.

Dorfleitner, G. , Utz, S. , and Wimmer, M. (2013). Where and when does it pay to be good? A global long-term analysis of esg investing. In *26th Australasian Finance and Banking Conference* .

Fan, M. , Kearney, F. , Li, Y. , and Liu, J. (2020). Momentum and the cross-section of stock volatility. *Available at SSRN 3977553*.

Freedman, M. and Stagliano, A. (1991). Differences in social-cost disclosures: a market test of investor reactions. Accounting, Auditing & Accountability Journal, 4(1).

Hanicova, D. and Vojtko, R. (2020). Backtesting ESG Factor Investing Strategies. *Available at SSRN 3633303*.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. The Journal of Finance, 48(1):65–91.

Jondeau, E. , Mojon, B. , and Pereira da Silva, L. A. (2021). Building benchmarks portfolios with decreasing carbon footprints. *Swiss Finance Institute Research Paper*, pages 21–91.

Kaiser, L. (2020). ESG integration: value, growth and momentum. Journal of Asset Management, 21(1):32–51.

Nagy, Z. , Kassam, A. , and Lee, L.-E. (2016). Can ESG add alpha? An analysis of ESG tilt and momentum strategies. The Journal of Investing, 25(2):113–124.

Nofsinger, J. R. and Varma, A. (2012). Individuals and their local utility stocks: preference for the familiar. Financial Review, 47(3):423–443.

Padysak, M. (n.d.). ESG Scores and Price Momentum Are More Than Compatible. *Available at SSRN:* https://ssrn.com/abstract=3650163 or http://dx.doi.org/10.2139/ssrn.3650163.

Verheyden, T. , Eccles, R. G. , and Feiner, A. (2016). ESG for all? The impact of ESG screening on return, risk, and diversification. Journal of Applied Corporate Finance, 28(2):47–55.